DOCUMENT RESUME

ED 342 808 TM 017 985

AUTHOR	Breland, Hunter M.; And Others
TITLE	A Study of Gender and Performance on Advanced Placement History Examinations. College Board Report No. 91-4.
INSTITUTION	College Entrance Examination Board, New York, N.Y.
REPORT NO	ETS-RR-91-61
PUB DATE	91
NOTE	44p.
AVAILABLE FROM	College Board Publications, Box 886, New York, NY 10101-0886 ($12.00).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Advanced Placement; Aptitude Tests; College Entrance Examinations; Comparative Testing; Estimation (Mathematics); *European History; High Schools; *High School Students; Multiple Choice Tests; Predictor Variables; Regression (Statistics); Scoring; *Sex Differences; *Student Placement; Test Bias; *United States History; Writing (Composition)
IDENTIFIERS	*Advanced Placement Examinations (CEEB); Free Response Test Items; Scholastic Aptitude Test

ABSTRACT

Several studies have shown that, on average, women perform slightly better than men on free-response tests, while men perform slightly better on multiple-choice tests. Two advanced placement examinations, United States History (USH) and European History (EH), were chosen for study because previous studies have shown that sex differences on the free-response parts of these tests were non-significant, while sex differences on the multiple-choice parts were large. Several ratings and analyses were made of random samples of 400 free-response booklets from the 1986 administrations of both examinations. Samples of 199 males and 199 females for USH and 198 males and 199 females for EH were used. The following significant predictors were identified: (1) the advanced placement multiple-choice score; (2) historical content; (3) English composition quality; and (4) the number of words written. Regression equations were developed to estimate overprediction or underprediction by sex of free-response scores from the multiple-choice score alone and from multiple predictors. Sampled cases were matched with Scholastic Aptitude Test and English Composition scores. Results suggest that format effects are real and cannot be attributed to bias in scoring or to totally irrelevant variables. When scoring was conducted analytically with a focus on historical content, no sex differences were observed in the free-response portions. Males and females differ in how they respond to two different but legitimate ways of assessing history skills, raising the question of how much influence basic skills and aptitude should have on outcomes of assessment. Twenty tables present study data, and 7 appendices contain supplemental information, including 14 tables and 15 test rating forms. An 18-item list of references is included. (SLD)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

ROBERT G. CAMERON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

College Board Report

No. 91-4

CG 022808

A Study of Gender and Performance on Advanced Placement History Examinations

Hunter M. Breland

with

Despina O. Danos
Helen D. Kahn
Melvin Y. Kubota
Marilyn W. Sudlow

BEST COPY AVAILABLE

A Study of Gender and Performance on Advanced Placement History Examinations

Hunter M. Breland

with

Despina O. Danos
Helen D. Kahn
Melvin Y. Kubota
Marilyn W. Sudlow

College Board Report No. 91-4
ETS RR No. 91-61

College Entrance Examination Board, New York, 1991

Hunter M. Breland is a Senior Research Scientist at ETS.
Despina O. Danos is a Senior Examiner at ETS.
Helen D. Kahn is an Examiner at ETS.
Melvin Y. Kubota is an Examiner at ETS.
Marilyn W. Sudlow is a Senior Examiner at ETS.

Acknowledgments

The author expresses appreciation to Randy Bennett, Brent Bridgeman, John Mazzeo, Robert Mislevy, and Lawrence Stricker for numerous helpful suggestions on earlier drafts of this report.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,800 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101–0886. The price is $12.

Copyright © 1991 by College Entrance Examination Board. All rights reserved.
College Board, Advance Placement Program, and the acorn logo are registered trademarks of the College Entrance Examination Board. AP is a trademark owned by the College Entrance Examination Board.

Printed in the United States of America.

CONTENTS

Abstract 1

Introduction 1

Method 2
- Sampling 2
- Ratings 3
- Variable Development 3
- Analytical Methods 4

Results 4
- Results for the Full Sample 4
- Results for the Historical Content Sample 8
- Results for the Matched ATP Sample 11

Discussion 13

Further Research Possibilities 15

References 15

Tables

1. Means, Standard Deviations, and Effect Sizes for Full Sample 5
2. Multiple Regression Predictions of AP Free-Response Scores Excluding the AP Objective Score, Full Sample 5
3. Multiple Regression Predictions of AP Part A Free-Response Scores Excluding the AP Objective Score, Full Sample 6
4. Multiple Regression Predictions of AP Part B Free-Response Scores Excluding the AP Objective Score, Full Sample 6
5. Multiple Regression Predictions of AP Free-Response Scores, Full Sample 6
6. Multiple Regression Predictions of AP Part A Free-Response Scores, Full Sample 7
7. Multiple Regression Predictions of AP Part B Free-Response Scores, Full Sample 7
8. Over/Underprediction of AP Free-Response Scores, Full Sample 7
9. Means, Standard Deviations, and Effect Sizes for Historical Content Sample 8
10. Multiple Regression Predictions of AP Free-Response Scores Excluding AP Objective Score, Historical Content Sample 9
11. Multiple Regression Predictions of AP Part A and Part B Free-Response Scores for U.S. History Excluding AP Objective Score, Historical Content Sample 9
12. Multiple Regression Predictions of AP Part A and Part B Free-Response Scores for European History Excluding AP Objective Score, Historical Content Sample 10

13. Multiple Regression Predictions of AP Free-Response Scores Using AP Objective Score and Ratings 10

14. Multiple Regression Predictions of AP Part A and Part B Free-Response Scores for U.S. History, Historical Content Sample 11

15. Multiple Regression Predictions of AP Part A and Part B Free-Response Scores for European History, Historical Content Sample 11

16. Over/Underprediction of AP Free-Response Scores, Historical Content Sample 12

17. Comparison of Predictions of AP Free-Response Scores When Predictor Variables Are Categorized with Respect to Probable Construct Relevance 12

18. Means, Standard Deviations, and Effect Sizes for Cases Matched with ATP Data 13

19. Multiple Regression Predictions of AP Free-Response Scores Using ATP Matched Variables for U.S. History 13

20. Multiple Regression Predictions of AP Free-Response Scores Using ATP Matched Variables for European History 14

Appendix A. English Rating Procedures 17

Appendix B. Handwriting, Neatness, and Length Estimation Procedures 17

Appendix C. U.S. History Content Rating Procedures 18

Appendix D. European History Content Rating Procedures 27

Appendix E. Correlation Tables 35

Appendix F. Reliability Estimates 38

Appendix G. Interaction Analyses 38

ABSTRACT

Several studies have shown that, on average, women perform slightly better than men on free-response tests, while men perform slightly better on multiple-choice tests. Studies of the College Board Advanced Placement (AP) Examinations have revealed a similar phenomenon. For almost all AP Examinations, men average better on both parts of the tests, but sex differences on the free-response parts are almost always less, and for some tests they are nonsignificant.

Two AP Examinations, U.S. History and European History, were selected for study because sex differences on the free-response parts were nonsignificant while sex differences on the multiple-choice parts were large. Random samples of free-response booklets were drawn from the 1986 administrations of both examinations. A number of ratings and analyses were made of the responses: English composition quality, historical content, responsiveness, factual errors, handwriting quality, neatness, and number of words written. All variables were then used to predict the free-response scores. Several significant predictors were observed: the AP multiple-choice score, historical content, English composition quality, and the number of words written.

Predictor variables were classified with respect to their degree of relevance to the measurement construct. The historical content ratings and multiple-choice history scores were considered to be directly relevant, English composition quality and words written to be indirectly relevant, and neatness and handwriting to be irrelevant. For the two free-response scores combined, the indirectly relevant variables, in combination, predicted about as well as the directly relevant variables. For the Part A free response, the indirectly relevant variables predicted best, while for the Part B free response, the directly relevant variables predicted best. Handwriting and neatness did not predict free-response scores well at all.

The regression equations developed were also used to estimate over- or underprediction by sex of free-response scores from the AP multiple-choice score alone and from multiple predictors. As expected, the AP multiple-choice score underpredicted female performance on the free-response part of the test and overpredicted male performance. For the AP U.S. History Examination, the over/underprediction effect size was reduced to a nonsignificant level when the English composition quality of the free responses was added to the predictor set. For European History, the over/underprediction effect size was also reduced to a nonsignificant level by the addition of English composition quality, although the reduction of the effect size was not as dramatic as that for U.S. History. No other predictor variables contributed importantly to the reduction of effect sizes.

To develop further insight into this test format phenomenon, AP files for the sampled cases were matched with College Board Admissions Testing Program (ATP) data to obtain Scholastic Aptitude Test (SAT) scores and English Composition Test scores. In U.S. History, the SAT-verbal, Test of Standard Written English (TSWE), and English Composition Test (ECT) scores made significant independent contributions to the prediction of AP free-response scores over and above that possible using the AP objective (multiple-choice) score, especially for Part A. In European History, SAT-verbal scores made significant independent contributions, although TSWE and ECT scores did not.

This study suggests that format effects are real and cannot be attributed to bias in scoring or to totally irrelevant variables. When scoring was conducted analytically with a focus on historical content, no sex differences were observed in the free-response portions. This is the same result observed for the regular administration readings, which are graded holistically and by readers different from those used for this study. The sexes appear to differ in how they respond to two legitimate but different ways of assessing history skills. The study raises the question of how much influence basic skills and aptitude should have on the outcomes of assessments in specific areas of achievement.

INTRODUCTION

Recent analyses of data from the AP Examinations reveal differences in male and female performance that have not been explained. Although males typically average higher scores than females on the objective (multiple-choice) portions of the examinations, the same difference is not observed in the free-response portions. Mazzeo, Schmitt, and Bleistein (1990) analyzed several AP Examinations and showed that, more often than not, the gender differences exhibited on the objective portions of the examinations were greater than those on the free-response portions. In most cases, males performed better than females on both portions, but in some cases the differences on the free-response portions were trivial. After examining possible reasons for these gender differences, Mazzeo et al. concluded that they were not the result of the generally low reliabilities of the free-response examinations or of differential item performance on the objective portions of the examinations. It was hypothesized that the observed differences may be due to construct-irrelevant factors such as penmanship, neatness, and the like, or to construct-relevant factors representing real differences between the sexes in academic competencies.

Past studies of free-response tests have also revealed gender differences, usually favoring females. Breland (1977), in a comparison of group differences in performance on the TSWE and essay scores written in first-year English courses, observed that females often performed better on essays than their TSWE scores would suggest. Breland and Griswold (1982) studied gender differences in basic skills measures for large samples of males and females who had entered one of the California State Universities and Col-

leges in 1977. The TSWE was used to predict performance on a 45-minute essay examination written as part of the California English Placement Test (EPT). The results showed that women generally wrote better essays than would be predicted by the TSWE.

Petersen and Livingston (1982) examined group differences for the College Board ECT, which includes for some administrations both an objective portion and an essay sample. In this study, females averaged higher than males on both the multiple-choice and free-response portions of the test, though the differences were smaller for the multiple-choice portion.

Klein (1989) studied gender differences in relation to the California Bar Examination, which includes a multiple-choice test, six one-hour essay questions, and two three-hour written performance problems. The results of this study showed that males performed better on the multiple-choice tests but females performed better on both the essay test and the written performance test.

Recently, Bridgeman and Lewis (1990) studied AP scores as predictors of college performance. The results of this study suggest that the multiple-choice portions of AP Examinations predict freshman grade-point average (GPA) better than the essay portions. But when AP history examination scores were used to predict performance in college history courses, the advantage of multiple-choice scores in prediction was muted. In only one of three comparisons (for European history) did the multiple-choice test predict better than the essay test. There was little difference between the sexes in average history course grades, and this was mirrored by a lack of differences between the sexes in AP essay grades, even though the AP multiple-choice scores favored males.

The observation that females perform better on free-response tasks is not uncommon in psychology. Sincoff and Sternberg (1988), in a paper on verbal fluency differences in school-age children, report that girls—especially those above age 11—tend to score higher than boys on verbal fluency tasks. They cite previous studies or reviews in support of this observation (Droege 1967; Maccoby and Jacklin 1974; Stevenson, Klein, Hale, and Miller 1968). In their own study, Sincoff and Sternberg found that girls improved more in verbal fluency with age than did boys. Sixth-grade girls received higher verbal fluency task scores than did sixth-grade boys.

In a summary paper, Sincoff and Sternberg (1987) distinguish two types of verbal ability: verbal comprehension and verbal fluency. Verbal comprehension is required primarily for reading and listening tasks, while verbal fluency is required primarily for writing and speaking tasks. It is with respect to verbal fluency, not verbal comprehension, that research shows females to have a slight advantage. The AP free-response examinations would appear to require more verbal fluency than verbal comprehension, even though parts of the tests require the reading of documents prior to writing an essay.

The AP U.S. History and European History Examinations that have been studied for this report have two free-response parts. Part A is a document-based question (DBQ) for which the examinee is provided several brief reading pieces or illustrations (paintings or cartoons, for example). Time is allowed for study of these documents, and questions are then posed that the examinee must answer using the documents. The DBQ is scored slightly differently in the two examinations. In U.S. History, additional credit is given for introducing outside information not provided in the documents; in European History, no additional credit is given for using outside information. Part B of the AP history examinations consists of several questions from which the examinee must choose one. For both tests, the free-response Parts A and B are scored holistically by readers, and the free-response scores are then combined with the objective (multiple-choice) test scores to produce the AP grade for the examination.

METHOD

The AP U.S. History and European History Examinations were chosen for study because they both showed significant differences between the sexes in multiple-choice scores (favoring males) accompanied by statistically nonsignificant differences in free-response scores. More specifically, it was decided to examine the 1986 data for these examinations as well as actual protocols of free responses that resulted in the scores. The 1986 examinations were chosen because they were to be destroyed as part of routine AP data-retention policies, so samples of the free-response booklets could easily be obtained and retained for research purposes.

Sampling

For each of the two history examinations, random samples of 400 free-response booklets were obtained from boxes in storage by means of random numbers generated to identify individual boxes and the positions of essays filed within each box (inches from the front of the box). Since more males than females took the examinations in 1986 (for U.S. History, 35,636 males and 31,427 females; for European History, 10,308 males and 7,698 females), random sampling was conducted until the response booklets of 200 males were obtained for each examination. Random sampling was continued until the response booklets of 200 females had also been obtained for each examination, rejecting any male samples beyond the first 200. The essays sampled were then matched with AP Examination grade data, and matches were obtained for 398 AP U.S. History Examinations and 397 AP European History Examinations. The free-response booklets were next subjected to a number of rating or estimation procedures. A second random sampling down to 300 (150 male and 150 female) cases for each examination was conducted to reduce the workload involved in some of the more complex rating procedures. A

third sample was developed through a match of cases with the College Board ATP data. This match resulted in samples of 276 cases for U.S. History and 272 cases for European History.

Ratings

Ratings or estimations were made to develop the following variables: handwriting quality, neatness, holistic English quality, mechanical English quality, historical content (main points), historical content (supporting evidence), historical content (factual errors), responsiveness to questions asked, and estimated number of words written.

The handwriting quality and neatness ratings and the estimate of words written were made by staff, the English ratings were made by English teachers who were experienced readers of English composition examinations, and the historical content ratings were made by graduate students of history.

History specialists at Educational Testing Service and an outside consultant assisted in the development of scoring guides for each free-response question in each of the two history examinations. There were a total of 13 different questions asked in the two examinations, and one question allowed for a choice of one out of three topics. Accordingly, a total of 15 different scoring guides were developed. These guides were then used by two graduate history students (one male and one female) to obtain ratings for 300 randomly sampled papers from the total available. For each free-response booklet, each graduate student rated one of the two free responses. (The first and second free responses in the booklets were randomly assigned such that each student read the same number of first and second responses.) As in the regular AP administration readings, the two scores were then summed to develop a free-response score for each sampled booklet. Thus, although the rating procedures and the actual raters were different, the historical content ratings were conducted such that they paralleled AP administration scoring. The historical content ratings were analytic in nature, while the AP administration scoring is essentially holistic.

Further details on these ratings and estimations are given in Appendixes A (English), B (Handwriting, Neatness, Length), C (U.S. History), and D (European History). The ratings and estimations resulted in the following initial scores:

- *Handwriting Quality.* This rating was made on a five-point scale by two different raters, one male and one female, working independently. No distinction was made between Part A and Part B handwriting quality, but a global rating was made for the entire test booklet.
- *Neatness.* This rating was made on a five-point scale by two different raters, one male and one female, working independently. As for handwriting quality, no distinction was made between Part A and Part B, but a global rating of neatness was made for the entire test booklet.
- *Holistic English.* This rating was made on a six-point scale by experienced scorers of written English examinations. These readers were not provided with the history questions asked, but were instructed to read the essays as if they were English examinations and to score them as they would normally score English examinations. Two different readers working independently rated each free-response booklet, making separate ratings for Part A and Part B. Half of the readers were male and half were female.
- *English Mechanics.* This rating was made on a five-point scale at the same time as the holistic English rating. Two different readers working independently rated each essay. Half of the readers were male and half were female.
- *Words Written.* An estimate of the number of words written for each of the two essays in each booklet was made by counting the lines written and the average number of words per line.
- *Main Points.* Possible main points that a response could contain were developed by history specialists experienced with the particular questions used in 1986. They took into account previously obtained free responses for these questions and their own judgments of what points might be made even if no previous responses contained them. Allowance was also made for original main points that a response might contain by including a space on the rating guide for "other main point."
- *Supporting Evidence.* The supporting evidence scoring guide was developed in the same way as the guide for main points.
- *Factual Errors.* A space was provided on the form for tallying the number of factual errors contained in a response based on the judgment of the rater.
- *Responsiveness.* This rating was made on a four-point scale and was intended mainly to identify responses that were not serious or that did not address the questions posed. Several blank booklets were included in the random sample, and these provided the minimum responsiveness rating. (This rating was discarded, however, after observing its close relationship with the historical content ratings.)

Variable Development

Some variables were obtained from the match with the AP data base, and others were developed from combinations of the ratings made. The following variables were obtained either from the data base match or by combining ratings:

- *AP Objective Score.* This score is derived from a formula-scored 100-item multiple-choice test. Examinees are allowed 75 minutes to answer these 100 questions. For the 398 U.S. History cases analyzed,

formula scores ranged from a high of 88 to a low of 4. For the 397 European History cases analyzed, formula scores ranged from a high of 85 to a low of 6.

- *AP Free-Response Score*. This score is developed in the AP procedures by summing each of the two scores obtained from the two responses in the test booklet. Each of the two essays is scored by a different reader on a 0 to 15 point scale and then summed to produce scores from 0 to 30. For the examinations sampled, the score range was 0 to 30.
- *Historical Content*. This score was developed by summing the main points and supporting evidence scores and then subtracting the number of factual errors. These scores ranged from 0 to 31 in U.S. History and from 0 to 30 in European History.
- *English Score*. This score was developed by combining the holistic English score and the English mechanics score. The holistic scores for the two essays in each booklet were summed and the English mechanics score, which consisted of a single rating for the entire test booklet, was then doubled and added to the sum of the two holistic scores to provide an approximately equal balance for each score.

Analytical Methods

Analyses were conducted separately for each of the three samples. The first sample analyzed was the full sample of 398 cases in U.S. History and 397 cases in European History. The second set of analyses was conducted using the random samples of 300 cases from each examination for which historical content ratings were conducted. The third set of analyses was based on reduced samples of cases for which a match with ATP data was possible. The ATP data allowed for analyses of SAT, TSWE, and ECT data combined with the AP data.

Intercorrelations were computed and regression analyses were conducted to predict AP free-response scores from available variables. Using the AP objective score as a base predictor, a residualized criterion was developed by subtracting the actual AP free-response score from the score predicted by the AP objective score. This residualized criterion represented overprediction if positive and underprediction if negative. Hierarchical regression techniques (Cohen and Cohen 1983) were used to determine the contribution of covariates to the over/underprediction for males and females and to examine interactions between sex and other variables. Statistical as well as practical significance of effect sizes was considered. Cohen's (1988) method was used to interpret the practical significance of effect sizes.

RESULTS

Results are reported separately for the three different samples analyzed: the full sample, the sample with historical content ratings, and the ATP matched sample.

Results for the Full Sample

The sampled cases resulted in the same general pattern of scores that had been observed in previous analyses of the 1986 (and other) AP history examinations. Table 1 shows large differences (effect size = .37) between males and females on objective scores for both history examinations but generally small differences for the AP free-response scores. For both history examinations, males are favored on the objective score means. Similar differences of .35 standard deviations were observed for the total of 35,636 males and 31,427 females who took the AP U.S. History Examination in 1986, and .47 standard deviations were observed for the 10,308 males and 7,698 females who took the AP European History Examination in 1986. In neither the sample of cases analyzed for this study nor the much larger samples of students who took the examinations in 1986 were statistically significant gender differences observed for the free-response portion of the tests, for which both free-response parts were summed.

Table 1 shows a difference, however, when Part A (DBQ) and Part B are considered separately. For Part A of the AP U.S. History Examination, the standardized difference in mean scores ($-.26$) is both statistically and practically significant. The negative sign for this difference indicates that females averaged higher scores. Statistically and practically significant differences favoring females are also observed in Table 1 for the U.S. History English score, holistic English score, Part A holistic English score, English mechanics score, handwriting rating, and neatness rating.

In sharp contrast, Table 1 indicates only two statistically significant male-female differences for European History: for the AP objective score and the handwriting rating.

Correlations among the variables for the two history examinations are given in Appendix E. In interpreting these correlations, it is important to be aware of the relatively low reliabilities of the rating variables as well as the experimental dependence among them. Low reliabilities attenuate correlations, and experimental dependence artificially inflates correlations (reliability estimates for variables are given in Appendix F). Despite these interpretational problems, some interesting observations are possible. First, relatively high correlations between the AP free-response score and the English score (.50 for U.S. History and .53 for European History) are noteworthy. Second, the correlations between handwriting and neatness and the AP free-response score are relatively low, indicating that these variables are not important influences on AP free-response scores. Third, words written correlates highly with the AP free-response score (.57 for U.S. History and .61 for European History).

Tables 2 through 7 present a series of multiple regression analyses in which the variables developed for the present study were used to predict the AP free-response scores. Table 2 presents an analysis in which the AP free-response score (Parts A and B combined) is predicted using the English score and words written. For this analysis, the AP objective score was excluded. Neither the handwriting rating nor the neatness rating contributed significantly in these

Table 1. Means, Standard Deviations, and Effect Sizes for Full Sample

Variable	Male		Female		Effect Size††
	Mean	S.D.	Mean	S.D.	
			U.S. History (N = *398*)†		
AP objective score	46.39	16.39	40.32	16.12	.37**
AP free-response score	11.86	3.91	12.25	3.88	−.10
Part A	6.09	2.25	6.68	2.22	−.26**
Part B	5.77	2.48	5.56	2.56	.08
English score	27.18	6.47	29.12	7.44	−.28**
Holistic English score	14.51	3.75	15.67	5.26	−.26**
Part A	7.33	2.24	7.95	1.87	−.30**
Part B	7.18	1.94	7.47	1.97	−.15
Words written	873.6	292.4	912.2	270.7	−.14
Part A	489.7	163.4	526.5	161.1	−.23*
Part B	383.9	163.7	385.7	156.4	−.01
English mechanics score	6.34	1.56	6.72	1.63	−.24*
Handwriting rating	2.86	.89	3.58	.71	−.90**
Neatness rating	2.98	1.11	3.31	.94	−.32**
			European History (N = *397*)†		
AP objective score	43.95	16.11	38.48	13.68	.37**
AP free-response score	12.56	4.06	12.63	3.86	−.02
Part A	7.24	2.41	7.15	2.32	.04
Part B	5.32	2.48	5.49	2.51	−.07
English score	28.26	7.20	28.78	6.26	−.08
Holistic English score	15.15	4.14	15.36	3.53	−.05
Part A	7.86	1.99	7.92	1.81	−.03
Part B	7.28	2.37	7.43	1.93	−.07
Words written	788.8	274.7	758.8	244.4	.11
Part A	448.5	154.1	430.1	160.3	.12
Part B	340.3	160.2	328.7	131.7	.08
English mechanics score	6.56	1.66	6.71	1.49	−.09
Handwriting rating	3.75	.78	4.00	.70	−.36**
Neatness rating	3.54	.76	3.65	.68	−.15

†For U.S. History, there were 199 males and 199 females; for European History, there were 198 males and 199 females.
††Effect size is defined as the difference between the means for each sex divided by the average of the two standard deviations.
*$p < .05$ (negative effect size indicates female means were higher)
**$p < .01$

Table 2. Multiple Regression Predictions of AP Free-Response Scores Excluding the AP Objective Score, Full Sample

Group/Predictors	R	b	beta	p
		U.S. History		
Total (N = *398*)				
English score	.63	.17658	.32	.0000
Words written		.00594	.43	.0000
Male (N = *199*)				
English score	.66	.18401	.30	.0000
Words written		.00621	.46	.0000
Female (N = *199*)				
English score	.60	.17423	.33	.0000
Words written		.00561	.39	.0000
		European History		
Total (N = *397*)				
English score	.66	.19707	.33	.0000
Words written		.00625	.41	.0000
Male (N = *198*)				
English score	.69	.19742	.35	.0000
Words written		.00622	.42	.0000
Female (N = *199*)				
English score	.63	.19417	.32	.0000
Words written		.00635	.40	.0000

predictions. What is interesting about Table 2 is that it shows that, for both U.S. History and European History, reasonably good predictions of free-response scores can be made from the quality of the English and the number of words written. The multiple correlations range from a low of .60 (for females in U.S. History) to a high of .69 (for males in European History). The beta weights indicate that words written is relatively more important than English quality. Obviously, however, words written is also closely associated with historical content. Analyses of historical content are presented below in the next section of results.

In Tables 3 and 4, an analysis similar to that of Table 2 is repeated for free-response Parts A and B separately. Again, the AP objective score has been excluded. Despite the low (and unknown) reliabilities of the Part A and Part B free-response scores, the multiple correlations still range from a low of .49 to a high of .64. Tables 3 and 4 suggest that for both history examinations, and for both Parts A and B, the writing quality and the length of the response are important determinants of the free-response scores assigned.

Tables 5 through 7 introduce the AP objective score to

the multiple regressions. The increase in multiple correlations (the range in Table 5 is from .66 to .72) suggests that a broad knowledge of history (as measured by the AP objective score) helps in responding to the free-response questions. However, the significant beta weights for the holistic English score and for words written indicate that these variables are still important for predicting AP free-response scores. For European History in Table 5, the neatness rating comes in as a significant contributor for females only. Of the variables making significant contributions to the prediction of the AP free-response score in Table 5, the number of words written was generally the most important, as indicated by the beta weights. This generalization holds true fairly well for Part A (see Table 6), and for Part B of U.S. History (see Table 7), but it does not hold for Part B of European History (see also Table 7).

Table 3. Multiple Regression Predictions of AP Part A Free-Response Scores Excluding the AP Objective Score, Full Sample

Group/Predictors	R	b	*beta*	p
		U.S. History		
Total (N = *398*)				
Holistic English, Part A	.57	.08189	.11	.0336
Words written, Part A		.00550	.40	.0000
English mechanics		.32472	.23	.0000
Male (N = *199*)				
Holistic English, Part A	.58	.14607	.14	.0650
Words written, Part A		.00606	.44	.0000
English mechanics		.17323	.13	.1044
Female (N = *199*)				
Holistic English, Part A	.55	.05177	.08	.2444
Words written, Part A		.00478	.35	.0000
English mechanics		.42701	.31	.0000
		European History		
Total (N = *397*)				
Holistic English, Part A	.60	.38479	.31	.0000
Words written, Part A		.00521	.35	.0000
Male (N = *198*)				
Holistic English, Part A	.64	.38469	.32	.0000
Words written, Part A		.00634	.40	.0000
Female (N = *199*)				
Holistic English, Part A	.55	.37762	.30	.0001
Words written, Part A		.00481	.33	.0000

Table 4. Multiple Regression Predictions of AP Part B Free-Response Scores Excluding the AP Objective Score, Full Sample

Group/Predictors	R	b	*beta*	p
		U.S. History		
Total (N = *398*)				
Holistic English, Part B	.59	.39889	.31	.0000
Words written, Part B		.00572	.36	.0000
Male (N = *199*)				
Holistic English, Part B	.60	.38502	.30	.0000
Words written, Part B		.00583	.38	.0000
Female (N = *199*)				
Holistic English, Part B	.59	.24060	.34	.0000
Words written, Part B		.00548	.33	.0000
		European History		
Total (N = *397*)				
Holistic English, Part B	.55	.40803	.35	.0000
Words written, Part B		.00433	.25	.0000
Male (N = *198*)				
Holistic English, Part B	.60	.41810	.40	.0000
Words written, Part B		.00392	.24	.0019
Female (N = *199*)				
Holistic English, Part B	.49	.39443	.30	.0001
Words written, Part B		.00494	.26	.0008

Table 5. Multiple Regression Predictions of AP Free-Response Scores, Full Sample

Group/Predictors	R	b	*beta*	p
		U.S. History		
Total (N = *398*)				
AP objective score	.71	.06995	.30	.0000
Holistic English score		.28583	.27	.0000
Words written		.06486	.35	.0000
Male (N = *199*)				
AP objective score	.71	.06268	.26	.0000
Holistic English score		.27317	.26	.0000
Words written		.00512	.38	.0000
Female (N = *199*)				
AP objective score	.70	.08474	.34	.0000
Holistic English score		.26002	.24	.0002
Words written		.00469	.33	.0000
		European History		
Total (N = *397*)				
AP objective score	.69	.061383	.24	.0000
Holistic English score		.139210	.24	.0000
Words written		.006059	.40	.0000
Neatness rating		.16694	.02	.1084
Male (N = *198*)				
AP objective score	.72	.062855	.25	.0000
Holistic English score		.13669	.24	.0005
Words written		.006284	.42	.0000
Neatness rating		.0020707	.00	.9878
Female (N = *199*)				
AP objective score	.66	.071193	.25	.0000
Holistic English score		.12509	.20	.0032
Words written		.005816	.37	.0000
Neatness rating		.37876	.13	.0200

Table 8 presents an analysis of over- and underprediction for males and females using hierarchical regression techniques (Cohen and Cohen 1983; Cohen 1988). The residual means are developed by subtracting the actual AP free-response score from the predicted AP free-response score. A positive residual indicates overprediction and a negative residual underprediction. The effect size (Cohen's *d*) is the standardized difference between the sexes in the prediction. Analyses of interactions between sex and predictor variables were conducted initially, and these analyses are included as Appendix G. No statistically significant interactions were detected.

As would be expected from the results shown in Table 1, Table 8 shows that the AP free-response score is overpredicted for males by the AP objective score and underpredicted for females. The effect size of .36 is both statistically

Table 6. Multiple Regression Predictions of AP Part A Free-Response Scores, Full Sample

Group/Predictors	R	b	*beta*	p
		U.S. History		
Total (N = 398)				
AP objective score	.62	.03475	.25	.0000
Holistic English score, Part A		.08648	.11	.0288
English mechanics score		.20994	.15	.0025
Words written, Part A		.00534	.39	.0000
Male (N = 199)				
AP objective score	.61	.03165	.23	.0000
Holistic English score, Part A		.16394	.16	.0329
English mechanics score		.06076	.04	.5704
Words written, Part A		.00509	.37	.0000
Female (N = 199)				
AP objective score	.63	.04614	.33	.0000
Holistic English score, Part A		.03577	.13	.3884
English mechanics score		.26426	.19	.0045
Words written, Part A		.00488	.35	.0000
		European History		
Total (N = 397)				
AP objective score	.61	.02109	.14	.0018
Holistic English score, Part A		.21976	.18	.0355
English mechanics score		.11273	.08	.3075
Words written, Part A		.00579	.38	.0000
Neatness rating		.10057	.06	.1403
Male (N = 198)				
AP objective score	.65	.02400	.16	.0071
Holistic English score, Part A		.20566	.17	.1513
English mechanics score		.12318	.08	.4279
Words written, Part A		.00681	.43	.0000
Neatness rating		.02168	.01	.8074
Female (N = 199)				
AP objective score	.56	.02091	.12	.0570
Holistic English score, Part A		.23905	.19	.1116
English mechanics score		.06643	.04	.6766
Words written, Part A		.00486	.32	.0000
Neatness rating		.20901	.14	.0508

Table 7. Multiple Regression Predictions of AP Part B Free-Response Scores, Full Sample

Group/Predictors	R	b	*beta*	p
		U.S. History		
Total (N = 398)				
AP objective score	.63	.03298	.22	.0000
Holistic English score, Part B		.43791	.34	.0000
English mechanics score		−.21054	−.13	.0331
Words written, Part B		.00497	.31	.0000
Male (N = 199)				
AP objective score	.62	.02898	.19	.0025
Holistic English score, Part B		.32681	.26	.0230
English mechanics score		−.04626	−.03	.7584
Words written, Part B		.00542	.36	.0000
Female (N = 199)				
AP objective score	.63	.03594	.23	.0006
Holistic English score, Part B		.51317	.40	.0001
English mechanics score		−.03269	−.02	.0148
Words written, Part B		.00463	.28	.0001
		European History		
Total (N = 397)				
AP objective score	.59	.03926	.24	.0000
Holistic English score, Part B		.38320	.33	.0002
English mechanics score		−.14994	−.09	.1799
Words written, Part B		.00393	.23	.0001
Neatness rating		.08644	.05	.2398
Male (N = 198)				
AP objective score	.64	.03666	.24	.0002
Holistic English score, Part B		.45386	.43	.0006
English mechanics score		−.23801	−.16	.0969
Words written, Part B		.00344	.22	.0069
Neatness rating		.00082	.00	.9929
Female (N = 199)				
AP objective score	.54	.05092	.28	.0000
Holistic English score, Part B		.26341	.20	.0986
English mechanics score		−.05142	−.03	.7688
Words written, Part B		.004653	.24	.0018
Neatness rating		.19415	.11	.1025

Table 8. Over/Underprediction of AP Free-Response Scores, Full Sample

Predictor Set	*Sex*	N	R	*Residual**		*Effect Size*	p
				Mean	*S.D.*		
			U.S. History				
1. AP objective score	M	199	.49	.55	3.44	.36	.00
	F	199		−.54	3.27		
2. AP objective score	M	199	.64	.20	3.00	.14	.17
English score	F	199		−.20	2.95		
3. AP objective score	M	199	.71	.18	2.74	.13	.19
English score	F	199		−.18	2.74		
Words written							
			European History				
1. AP objective score	M	198	.42	.34	3.61	.19	.06
	F	199		−.34	3.53		
2. AP objective score	M	198	.61	.13	3.09	.08	.39
English score	F	199		−.13	3.15		
3. AP objective score	M	198	.69	.25	2.78	.18	.08
English score	F	199		−.25	2.89		
Words written							

*Residual = predicted free-response score minus actual free-response score.

BEST COPY AVAILABLE

and practically significant (Cohen considers .20 a small effect size). For the second predictor set, however, the English score is introduced to the prediction, and the effect size is reduced to .14, which is not significant either statistically or practically. A small further reduction in the effect size is produced by the addition of words written in the third predictor set.

Table 8 indicates that effect size for European History is only *nearly* significant statistically (at the .05 level), even though males' scores are overpredicted and females' underpredicted by the AP objective score, as they were for U.S. History. The addition of the English score to the prediction reduces the already small effect size to only .08, which is clearly not significant statistically. Oddly, the addition of words written *increases* the effect size back to about where it was before the English score was added. This oddity may have resulted from the fact that males actually wrote more words for European History than did females, a situation opposite that observed for U.S. History.

Results for the Historical Content Sample

What we have termed the historical content sample was a random sample of 300 cases (150 male and 150 female) drawn from the full sample of 398 cases in U.S. History and 300 cases drawn from the 397 cases in European History. The sample size was reduced because of the labor-intensiveness of the historical content ratings. Table 9 presents the means, standard deviations, and effect sizes for this reduced sample, and the patterns are quite similar to those of Table 1 for the full sample. Of special interest in Table 9 is that the historical content score effect size is nonsignificant, as was that for the AP free-response score (in both this sample and the full sample). This indicates that the explanation for the sex differences between AP objective and free-response scores is not to be found in the free-response rating methods. Both holistic ratings (as in the regular administration ratings) and analytical ratings (as done for the present study) yield similar nonsignificant effect sizes.

Tables 10 through 13 report the results of regression analyses for the historical content sample when the AP objective score was excluded. Table 10 presents the results obtained when the AP free-response score (Parts A and B combined) was predicted from the historical content score, the English score, and the number of words written. Note that neither handwriting nor neatness contributed significantly in this prediction. The beta weights of Table 10 show that the

Table 9. Means, Standard Deviations, and Effect Sizes for Historical Content Sample

Variable	Male (N = 150)		Female (N = 150)		Effect Size
	Mean	S.D.	Mean	S.D.	
			U.S. History		
AP objective score	44.95	16.25	39.58	15.83	.33**
AP free-response score	11.75	3.56	11.93	3.78	−.05
Part A score	6.08	2.16	6.50	2.13	−.20
Part B score	5.67	2.22	5.43	2.53	.10
Historical content score	15.21	5.86	15.27	5.54	−.01
Main points	8.23	2.74	8.32	2.45	−.03
Supporting evidence	6.99	4.26	6.95	4.12	.01
Factual errors	.48	.76	.56	.97	−.09
English score	27.02	6.52	28.83	7.50	−.26*
Holistic English	14.43	3.75	15.53	5.66	−.23*
English mechanics	6.29	1.59	6.65	1.57	−.23*
Handwriting rating	2.87	.88	3.61	.72	−.92**
Neatness rating	3.06	1.06	3.35	.93	−.29**
Words written	853.7	278.7	886.0	266.4	−.12
			European History		
AP objective score	43.80	16.48	37.87	13.68	.39**
AP free-response score	12.43	4.23	12.61	4.04	−.04
Part A score	7.20	2.45	7.02	2.29	.08
Part B score	5.23	2.58	5.59	2.61	−.14
Historical content score	14.20	6.02	13.91	5.62	.05
Main points	6.31	2.61	6.07	2.30	.10
Supporting evidence	7.89	4.63	7.83	4.25	.01
Factual errors	.29	.55	.26	.51	.06
English score	28.42	7.42	28.15	6.34	.04
Holistic English	15.23	4.31	15.01	3.60	.06
English mechanics	6.59	1.68	6.57	1.49	.01
Handwriting rating	3.75	.75	4.01	.67	−.37**
Neatness rating	3.51	.62	3.61	.72	−.10
Words written	800.8	293.2	753.9	258.5	.17

*$p < .05$ (negative effect size indicates female means were higher)
**$p < .01$

BEST COPY AVAILABLE

Table 10. Multiple Regression Predictions of AP Free-Response Scores Excluding AP Objective Score, Historical Content Sample

Groups/Predictors	r	R*	*beta*	p
		U.S. History		
Total (N = 300)				
Historical content score	.63	.71 (.62)	.41	.0000
English score	.51		.23	.0000
Words written	.54		.24	.0000
Male (N = 150)				
Historical content score	.61	.70 (.64)	.40	.0000
English score	.53		.27	.0003
Words written	.57		.29	.0003
Female (N = 150)				
Historical content score	.66	.72 (.60)	.46	.0000
English score	.50		.27	.0023
Words written	.52		.28	.0013
		European History		
Total (N = 300)				
Historical content score	.69	.76 (.70)	.40	.0000
English score	.63		.23	.0000
Words written	.64		.24	.0000
Male (N = 150)				
Historical content score	.70	.77 (.74)	.35	.0000
English score	.67		.26	.0012
Words written	.68		.26	.0015
Female (N = 150)				
Historical content score	.68	.75 (.68)	.43	.0000
English score	.58		.22	.0020
Words written	.61		.25	.0005

*The *R* values in parentheses represent predictions excluding the historical content score.

historical content score was most heavily weighted in this prediction, but the English score and words written still made independent contributions to the prediction. The incremental contribution of the historical content score is indicated by the figures in parentheses under *R* in the table. These figures represent the multiple correlation *excluding* the historical content score; the difference between the two *R*'s represents the incremental contribution of the historical content score beyond the prediction possible using only the English score and words written. It is also of interest to compare the *R* values in parentheses, which represent the prediction possible using the English score and words written, with the zero-order *r*'s for the historical content score. These values are of about the same order of magnitude, suggesting that the AP free-response score is equally well predicted by ratings of historical content and by other variables representing English skills and verbal fluency.

Table 11 is similar to Table 10 except that results are presented for AP free-response Parts A and B separately. For Part A, the beta weights for the historical content score are not the largest weights, as they were for Parts A and B combined (except for females). Most important, however, is to note that the *R* values in parentheses, representing the prediction possible using the English score and words written for Part A, are for all samples larger than the zero-order *r*'s for the historical content score. For Part B, the values in parentheses are all less than the zero-order *r*'s for the historical content score, which also has the largest beta weight for all samples.

Table 12 presents the same kind of analysis for European History. In this analysis, the neatness rating enters the prediction in a small way and is statistically significant only for Part A, for the total sample, and for females. As for U.S. History, the beta weights for the historical content score are not dominant for Part A, and the *R* values in parentheses—here representing the prediction possible using the English score, words written, and neatness—are all larger than the zero-order *r* for the historical content score. Thus, the Part A results for European History are similar to the Part A results for U.S. History, as are the Part B results. The beta weights for the historical content score dominate the prediction, and the prediction using the historical content score alone (the zero-order *r*) is slightly larger than the *R*'s in parentheses, representing prediction without the historical content score.

Table 13 adds the AP objective score to the prediction. The multiple *R*'s are not much different from those of Table 11, showing that the AP objective score does not add much to the prediction of the AP free-response score. This outcome is in part a result of the experimental dependence between the other three predictors and the AP free-response score.

Table 11. Multiple Regression Predictions of AP Part A and Part B Free-Response Scores for U.S. History Excluding AP Objective Score, Historical Content Sample

Groups/Predictors	r	R*	*beta*	p
		Predicting Part A Score		
Total (N = 300)				
Historical content score	.44	.58 (.54)	.24	.0000
English score	.44		.23	.0000
Words written	.46		.29	.0000
Male (N = 150)				
Historical content score	.40	.56 (.54)	.18	.0174
English score	.42		.21	.0077
Words written	.50		.34	.0000
Female (N = 150)				
Historical content score	.48	.58 (.53)	.29	.0002
English score	.45		.24	.0015
Words written	.42		.24	.0009
		Predicting Part B Score		
Total (N = 300)				
Historical content score	.64	.69 (.58)	.43	.0000
English score	.43		.16	.0008
Words written	.53		.24	.0000
Male (N = 150)				
Historical content score	.65	.70 (.59)	.46	.0000
English score	.46		.18	.0065
Words written	.54		.21	.0035
Female (N = 150)				
Historical content score	.62	.69 (.57)	.45	.0000
English score	.46		.14	.0401
Words written	.54		.27	.0002

*The *R* values in parentheses represent predictions excluding the historical content score.

Table 12. Multiple Regression Predictions of AP Part A and Part B Free-Response Scores for European History Excluding AP Objective Score, Historical Content Sample

Groups/Predictors	r	R*	*beta*	p
		Predicting Part A Score		
Total (N = 300)				
Historical content score	.53	.68 (.65)	.24	.0000
English score	.57		.26	.0000
Words written	.57		.28	.0000
Neatness score	.24		.10	.0284
Male (N = 150)				
Historical content score	.57	.69 (.68)	.21	.0119
English score	.62		.32	.0001
Words written	.61		.28	.0010
Neatness score	.20		.05	.4407
Female (N = 150)				
Historical content score	.49	.63 (.61)	.27	.0006
English score	.52		.19	.0167
Words written	.54		.29	.0002
Neatness score	.30		.16	.0173
		Predicting Part B Score		
Total (N = 300)				
Historical content score	.59	.65 (.57)	.38	.0000
English score	.52		.20	.0007
Words written	.52		.18	.0023
Neatness score	.15		.00	.9510
Male (N = 150)				
Historical content score	.62	.66 (.60)	.37	.0000
English score	.56		.21	.0227
Words written	.57		.19	.0358
Neatness score	.07		−.04	.5596
Female (N = 150)				
Historical content score	.57	.64 (.55)	.39	.0000
English score	.48		.21	.0082
Words written	.49		.20	.0110
Neatness score	.22		.01	.9275

*The *R* values in parentheses represent predictions excluding the historical content score.

Table 14, for U.S. History, shows that for Part A, the historical content score does not dominate the prediction, but it does for Part B. Table 15, for European History, also shows that the historical content score tends to dominate the prediction for Part B but not for Part A.

Similar to Table 8, Table 16 is an analysis of over- and underprediction for males and females using hierarchical regression techniques. As for the full sample analysis of Table 8, residual means are developed by subtracting the actual AP free-response score from the predicted AP free-response score. Positive residuals indicate overprediction and negative residuals underprediction. The effect size is the standardized difference between the sexes. As in the earlier analyses, possible interactions between sex and predictor variables were examined initially, and no statistically significant interactions were detected.

The results presented in Table 16 are similar to those for the full sample (Table 8). The AP objective score underpredicts female free-response scores and overpredicts those for males. For U.S. History, when the English score is added to the prediction, the effect size becomes both statistically and practically insignificant. The addition of the historical content score to the prediction reduces the effect size slightly more. Adding words written did not reduce the effect size further. A less dramatic reduction of the effect size occurred for European History when the English score and the historical content score were added to the prediction. For this sample, over/underprediction analyses were not conducted separately for free-response Parts A and B.

The variables available for this sample can also be classified with respect to their degree of probable construct relevance. The AP objective score and the historical content score might be considered to be of direct construct relevance (DCR), since they are both clearly history scores and both represent specific areas of history. The English score and words written might be classified in an indirect construct relevance (ICR) category, since the former represents a skill that historians should have and the latter is related indirectly to points made and the supporting evidence provided in the

Table 13. Multiple Regression Predictions of AP Free-Response Scores Using AP Objective Score and Ratings

Groups/Predictors	r	R	*beta*	p
		U.S. History		
Total (N = 300)				
AP objective score	.50	.72	.20	.0001
Historical content score	.63		.29	.0000
English score	.51		.20	.0000
Words written	.54		.27	.0000
Male (N = 150)				
AP objective score	.49	.71	.18	.0125
Historical content score	.61		.25	.0020
English score	.53		.22	.0013
Words written	.57		.28	.0001
Female (N = 150)				
AP objective score	.53	.73	.22	.0035
Historical content score	.66		.37	.0001
English score	.50		.16	.0142
Words written	.52		.28	.0001
		European History		
Total (N = 300)				
AP objective score	.43	.77	.13	.0024
Historical content score	.69		.35	.0000
English score	.63		.18	.0006
Words written	.64		.25	.0000
Neatness score	.23		.08	.0359
Male (N = 150)				
AP objective score	.49	.79	.20	.0007
Historical content score	.70		.27	.0010
English score	.67		.21	.0071
Words written	.68		.29	.0003
Female (N = 150)				
AP objective score	.39	.75	.05	.2567
Historical content score	.68		.42	.0000
English score	.58		.22	.0031
Words written	.61		.27	.0004

Table 14. Multiple Regression Predictions of AP Part A and Part B Free-Response Scores for U.S. History, Historical Content Sample

Groups/Predictors	r	R	*beta*	p
	Predicting Part A Score			
Total (N = 300)				
AP objective score	.38	.61	.22	.0000
Historical content score	.44		.14	.0176
English score	.44		.25	.0002
Words written	.46		.31	.0000
Male (N = 150)				
AP objective score	.38	.58	.19	.0155
Historical content score	.40		.10	.2261
English score	.42		.19	.0136
Words written	.50		.33	.0000
Female (N = 150)				
AP objective score	.43	.63	.30	.0002
Historical content score	.48		.15	.0716
English score	.45		.17	.0249
Words written	.42		.32	.0000
	Predicting Part B Score			
Total (N = 300)				
AP objective score	.42	.70	.09	.0817
Historical content score	.64		.39	.0000
English score	.43		.14	.0025
Words written	.53		.24	.0000
Male (N = 150)				
AP objective score	.42	.69	.09	.1951
Historical content score	.65		.43	.0000
English score	.46		.16	.0262
Words written	.54		.21	.0032
Female (N = 150)				
AP objective score	.42	.69	.06	.4121
Historical content score	.62		.41	.0000
English score	.43		.13	.0525
Words written	.54		.27	.0002

Table 15. Multiple Regression Predictions of AP Part A and Part B Free-Response Scores for European History, Historical Content Sample

Groups/Predictors	r	R	*beta*	p
	Predicting Part A Score			
Total (N = 300)				
AP objective score	.27	.68	.06	.1828
Historical content score	.53		.22	.0000
English score	.57		.24	.0000
Words written	.57		.29	.0000
Neatness score	.24		.11	.0187
Male (N = 150)				
AP objective score	.30	.70	.10	.1305
Historical content score	.57		.18	.0364
English score	.62		.29	.0008
Words written	.61		.30	.0004
Neatness score	.20		.05	.3861
Female (N = 150)				
AP objective score	.24	.64	.03	.6987
Historical content score	.49		.26	.0006
English score	.52		.18	.0267
Words written	.54		.29	.0002
Neatness score	.30		.17	.0163
	Predicting Part B Score			
Total (N = 300)				
AP objective score	.44	.67	.18	.0005
Historical content score	.59		.33	.0000
English score	.52		.16	.0105
Words written	.52		.18	.0018
Neatness score	.15		.03	.5541
Male (N = 150)				
AP objective score	.52	.70	.25	.0003
Historical content score	.62		.30	.0004
English score	.56		.14	.1209
Words written	.57		.19	.0305
Neatness score	.07		−.02	.8001
Female (N = 150)				
AP objective score	.52	.65	.15	.0438
Historical content score	.62		.32	.0001
English score	.56		.17	.0397
Words written	.57		.21	.0079
Neatness score	.07		.05	.5215

response. Handwriting quality and neatness can be classified in a third category, no construct relevance (NCR), since it is difficult to argue that these are necessary for a historian (even though they might be desirable).

Table 17 compares the prediction of AP free-response scores using these different categories of variables. When free-response Parts A and B are combined, the DCR variables appear to be slightly better predictors than the ICR variables in terms of multiple correlations, while the NCR variables tend not to predict the free-response score at all. For Part A, however, the ICR variables appear to predict better than the DCR variables, and for Part B, the DCR variables appear to predict better than the ICR variables. This result suggests that the Part B free responses are more construct relevant than the Part A free responses.

Results for the Matched ATP Sample

To further explain the differences between the scores of males and females on the multiple-choice and free-response portions of the two AP Examinations, additional analyses were conducted using a match of College Board AP data and College Board ATP data. These analyses allow for a more reliable comparison of writing skills and also make available aptitude score data for the students in the sample.

A total of 276 cases were matched with ATP program data for U.S. History and a total of 272 cases for European History. Table 18 shows a comparison of male and female mean scores for the matched variables. The AP objective and free-response score means have the same pattern as for the previously analyzed samples, with males scoring significantly higher on both objective portions and with no significant differences between males and females on the free-response portions. For U.S. History, males in this sample had significantly higher SAT-mathematical and SAT reading scores. Differences in ECT score means are substantial for both tests, favoring females, and these differences approach significance at the .05 level. The effect size of −.31

Table 16. Over/Underprediction of AP Free-Response Scores, Historical Content Sample

Predictor Set	Sex	N	R	Residual* Means	Residual* S.D.	Effect Size	p
			U.S. History				
1. AP objective score	M	150	.50	.39	3.10	.25	.03
	F	150		−.39	3.22		
2. AP objective score	M	150	.61	.13	2.80	.09	.44
English score	F	150		−.13	2.98		
3. AP objective score	M	150	.69	.05	2.61	.04	.75
English score	F	150		−.05	2.68		
Historical content score							
			European History				
1. AP objective score	M	150	.43	.43	3.69	.23	.04
	F	150		−.43	3.72		
2. AP objective score	M	150	.66	.31	2.96	.20	.08
English score	F	150		−.31	3.06		
3. AP objective score	M	150	.74	.24	2.72	.17	.13
English score	F	150		−.24	2.76		
Historical content score							

*Residual = predicted free-response score minus actual free-response score.

Table 17. Comparison of Predictions of AP Free-Response Scores When Predictor Variables Are Categorized with Respect to Probable Construct Relevance

Predictor Sets	Group	Multiple Correlation with: Part A	Part B	Parts A & B
	*U.S. History (N = 296)**			
Direct construct relevance				
AP objective score	Total	.48	.64	.65
Historical content score	Male	.44	.66	.63
	Female	.52	.62	.67
Indirect construct relevance				
English score	Total	.54	.58	.62
Words written	Male	.54	.59	.64
	Female	.53	.57	.60
No construct relevance				
Handwriting quality	Total	.08	.04	.04
Neatness	Male	.08	.07	.07
	Female	.07	.08	.02
	*European History (N = 289)**			
Direct construct relevance				
AP objective score	Total	.46	.57	.64
Historical content score	Male	.48	.59	.65
	Female	.43	.56	.64
Indirect construct relevance				
English score	Total	.56	.48	.62
Words written	Male	.59	.49	.64
	Female	.53	.49	.60
No construct relevance				
Handwriting quality	Total	.15	.07	.12
Neatness	Male	.18	.09	.14
	Female	.11	.10	.10

*In U.S. History, there were 148 males and 148 females; in European History, there were 142 males and 147 females. The number of cases is reduced for this comparison because cases with no response to either Part A or Part B were excluded.

for U.S. History is substantial. The lower number of matched ECT cases, however, makes statistical significance difficult to attain. Only 140 of the 276 U.S. History cases and only 167 of the 272 European History cases had ECT scores. The smaller differences for the TSWE reflect the lower difficulty of the TSWE as well as the truncation of scores at higher levels (TSWE scores are truncated at 60, meaning that all higher scores are reported at 60). For a high-ability group such as AP candidates, the variation in TSWE scores is limited by its lower difficulty and the truncation.

Table 19 presents the results of multiple regression analyses for U.S. History cases in which free-response scores were predicted using the matched ATP variables. These results show that, in addition to the AP objective score, TSWE, ECT, and SAT-verbal scores all contributed significantly in the prediction of U.S. History free-response scores. The best contribution was made when TSWE and ECT scores were added. Table 19 also shows that the TSWE and ECT contribute more to predictions of the Part A score than the Part B score.

Table 20 shows the results obtained from multiple regression analyses for European History cases. In none of these analyses did English test scores contribute to the prediction of free-response scores, but SAT-verbal scores contributed significantly. For the total AP free-response score (the sum of Parts A and B), the major contributor was the AP objective score, though statistically significant contributions were made by SAT-verbal scores. For Part A, SAT-verbal scores contributed about as much to the prediction as did the AP objective score. As for U.S. History, the SAT-verbal scores did not contribute significantly in the prediction of AP European History Part B scores.

Table 18. Means, Standard Deviations, and Effect Sizes for Cases Matched with ATP Data

Variable	Male		Female		Effect Size
	Mean	S.D.	Mean	S.D.	
			U.S. History (N = 276)†		
AP objective score	46.23	15.95	40.17	15.76	.38*
AP free-response score	12.25	3.62	12.33	3.58	−.02
Part A score	6.30	2.20	6.75	2.29	−.20
Part B score	5.95	2.34	5.58	2.29	.16
SAT-V score	53.70	8.28	51.73	8.78	.23
SAT-M score	59.83	9.06	53.80	9.25	.66*
SAT reading score	54.19	9.25	51.70	9.54	.26**
SAT vocabulary score	53.54	8.88	52.05	9.54	.16
TSWE score	50.73	7.34	51.27	7.30	−.07
ECT score (N = 140)	54.26	8.96	57.10	9.20	−.31
			European History (N = 272)††		
AP objective score	44.50	16.16	38.87	13.54	.38*
AP free-response score	12.54	4.20	12.39	3.91	.04
Part A score	7.13	2.46	7.12	2.42	.00
Part B score	5.40	2.46	5.27	2.41	.15
SAT-V score	55.31	9.23	54.34	8.42	.11
SAT-M score	59.38	9.69	55.18	9.70	.43*
SAT reading score	55.46	9.36	54.34	9.76	.12
SAT vocabulary score	55.34	10.26	54.62	9.49	.07
TSWE score	50.64	8.69	51.85	7.25	−.15
ECT score (N = 167)	55.76	10.03	57.17	8.79	−.15

†For U.S. History, there were 138 males and 138 females, except for the ECT score, where there were 72 males and 68 females.
††For European History, there were 136 males and 136 females, except for the ECT score, where there were 87 males and 80 females.
*$p < .01$ (negative effect size indicates female means were higher)
**$p < .05$

Table 19. Multiple Regression Predictions of AP Free-Response Scores Using ATP Matched Variables for U.S. History

Predictor Sets	N	b	*beta*	p	R
Predicting AP free-response score					
1. AP objective score	276	.09340	.42	.00	.54
TSWE		.10523	.22	.00	
2. AP objective score	276	.09427	.42	.00	.51
SAT-V		.05864	.14	.03	
3. AP objective score	140	.08204	.37	.00	.51
ECT		.09269	.25	.00	
Predicting Part A score					
1. AP objective score	276	.04368	.31	.00	.46
TSWE		.07386	.24	.00	
2. AP objective score	276	.04782	.34	.00	.41
SAT-V		.02996	.11	.10	
3. AP objective score	140	.03540	.24	.00	.44
ECT		.07374	.30	.00	
Predicting Part B score					
1. AP objective score	276	.04972	.35	.00	.39
TSWE		.03136	.10	.10	
2. AP objective score	276	.04645	.32	.00	.39
SAT-V		.02867	.11	.12	
3. AP objective score	140	.04664	.31	.00	.33
ECT		.01895	.08	.38	

DISCUSSION

One explanation for the observed differences between the sexes on scores for multiple-choice and free-response tests may be that these different formats assess different skill constructs. The AP multiple-choice history portions appear from their content to assess a broad range of historical knowledge. The AP free-response history portions appear from their content to assess a relatively narrower range of historical knowledge and to require some English composition skills and verbal aptitude. The free-response DBQs appear to require relatively more English skills and verbal aptitude and relatively less historical knowledge than the nondocument-based questions. The statistical analyses conducted for this investigation tend to support these content-based observations.

Analyses of total AP free-response scores (Parts A and B combined) showed, for both the U.S. History and European History examination forms studied, that four different variables contributed significantly to the prediction of total AP free-response scores: historical content, English skills, amount written, and the AP multiple-choice score. Historical content was determined independently of the judgments of AP Readers, and the English skills variables were based on judgments by experienced readers of English composi-

Table 20. Multiple Regression Predictions of AP Free-Response Scores Using ATP Matched Variables for European History

Predictor Sets	N	b	*beta*	p	R
Predicting AP free-response score					
1. AP objective score	272	.09452	.35	.00	.45
SAT-V		.06920	.15	.02	
2. AP objective score	272	.09532	.36	.00	.45
SAT reading score		.06694	.16	.01	
3. AP objective score	272	.09568	.36	.00	.45
SAT vocabulary score		.05394	.13	.04	
Predicting Part A score					
1. AP objective score	272	.03167	.20	.00	.34
SAT-V		.05245	.20	.00	
2. AP objective score	272	.03396	.21	.00	.33
SAT reading score		.04567	.18	.01	
3. AP objective score	272	.03427	.21	.00	.33
SAT vocabulary score		.04458	.18	.01	
Predicting Part B score					
1. AP objective score	272	.06284	.39	.00	.42
SAT-V		.01675	.06	.35	
2. AP objective score	272	.06136	.38	.00	.43
SAT reading score		.02126	.08	.20	
3. AP objective score	272	.06530	.41	.00	.42
SAT vocabulary score		.00937	.04	.55	

Neither TSWE nor ECT scores contributed significantly to the prediction of European History free-response scores; thus they do not appear in this table.

tions who were unaware of the historical questions to which the protocols responded. The number of words written in a given time period is an indicator of verbal fluency. The AP multiple-choice portion assesses a broad range of historical knowledge. Since all four of these variables made independent contributions to the prediction of the AP free-response score, it is suggested that all four were important influences on the scores assigned by AP Readers in the administrations studied.

Since the Part A (DBQ) and Part B free responses were based on different types of questions, it was of special interest to examine these parts separately. Several observations were made. First, it was noted that the historical content score contributed more to Part B than to Part A predictions for both examinations. Second, for Part A of both history examinations, the historical content score correlated less with the AP free-response score than did a combination of English skill and number of words written. For Part B of both examinations, the historical content score correlated better with the free-response score than did a combination of English skill and words written.

These Part A and Part B differences suggested a need to categorize predictor variables with respect to their degree of construct relevance. Two variables, the historical content score and the AP objective score, were classified as having direct construct relevance (DCR). This seemed a reasonable classification, since both are not only history scores but are also history scores for specific areas of history (U.S. or European). The English score and words written were classified as having indirect construct relevance (ICR) because it can be argued that historians need language skills and because the amount written is indirectly related to historical points made and to supporting evidence provided. Handwriting quality and neatness were classified as having no construct relevance (NCR), even though these may be desirable.

Multiple correlations between these classified variables and the AP free-response scores showed that the DCR and ICR variables predicted the AP free-response scores equally well when both free-response scores were summed. For both history examinations, however, the Part B responses were predicted best by the DCR variables and the Part A responses were predicted best by the ICR variables. AP free-response scores were not well predicted by the NCR variables, which is supportive of their construct validity. However, since the Part A (DBQ) free responses were predicted better by ICR than DCR variables, this indicates that DBQs may be less construct relevant than Part B questions.

The same indication occurred when ATP variables were related to the Part A and Part B scores. In these analyses, the AP objective score, TSWE, and ECT were used as predictors. To simplify the analyses, the AP objective score was paired with one of the other scores. In these analyses, the writing skill scores clearly contributed more to the predictions for Part A of U.S. History, and the aptitude test scores clearly contributed more to the predictions for Part A of European History. Surprisingly, the SAT-verbal score contributed as much to the prediction of the European History Part A score as did the AP objective score. For both tests, the AP objective score contributed more than either English composition scores or verbal ability scores in predictions of the Part B score.

Although the analyses conducted suggest that DBQs of the type used for the 1986 AP U.S. History and European History Examinations may be less construct relevant than the types of questions used for Part B, historians have a different view. From this perspective, a broad range of historical knowledge, skill in working with historical documents, and skill in generating historical arguments are all important for students of history. Especially important to historians are the primary documents used in the document-based part of the examinations. Holt (1990) has observed that:

> Primary documents can provide students a basis for developing fresh interpretations, for writing new narratives about historical problems. Alternative historical texts force them to make choices among competing narratives written by others. In either case they are challenged not only to think for themselves but to think creatively. (p. 42)

Messick's (1989) discussion of "construct underrepresentation" and "construct irrelevance" may be helpful here. From the historian's perspective, a multiple-choice test of the type used in AP Examinations might be viewed as underrepresenting what historians do, since this test does not require the student to work with historical documents,

primary or otherwise. Moreover, it does not encourage the use of primary documents in the teaching of history. On the other hand, from the measurement specialist's perspective, the document-based questions of the type used in the 1986 AP history examinations introduce what might be viewed as construct-irrelevant variance: reading and writing skills and verbal aptitude more generally.

To the degree that these more general skills are required, females may have a slight advantage. Even though males may have more knowledge of history, as suggested by the AP multiple-choice scores, females may have slightly more skill on average in English and verbal fluency, as suggested by past research. These skills may be the reason why differences between the sexes in AP free-response scores are less than differences between the sexes in AP objective scores.

FURTHER RESEARCH POSSIBILITIES

The study reported has obvious limitations that might be addressed in future studies. It was limited to an investigation of only two AP Examinations administered in only one year, 1986. Both of these were history examinations. Further research might investigate other examinations administered in other years.

A second limitation of the present study was that the dependent variable, the AP free-response score, was not independent of the various ratings made. The scores assigned by the AP Readers were based on the same protocols as the ratings conducted for this study. Consequently, when the ratings are used as covariates, one principle of experimental design is violated: the principle that covariates should be independent of treatment variables. This kind of independence is usually designed into studies by making sure that covariate measures are obtained prior to the collection of treatment data. It may be that an appropriate experimental design could be devised to examine the influence of English skill and verbal fluency on free-response scores.

A third limitation of the present study was that a more precise examination of the error structure in the rating variables was not possible. This problem begins with the fact that, in the AP readings, only one Reader reads both parts of the free-response booklet. The two grades are then combined into a single free-response grade. Although rough estimates can then be made of the reliability of the combined grade, no estimate at all is possible of the reliabilities of the two components. Multiple readings of each part would allow for better estimation of variance components and thus better estimates of reliability. If better reliability estimates were available for the AP free-response scores, it would then make sense to do the same for the experimental ratings. A thorough analysis of the error structure in all ratings would be essential for studying the construct relevance of free-response scores. This research would require AP Readers to assemble to conduct a second set of readings of some recent administration, multiple ratings of historical content by a separate set of Readers, multiple ratings of English composition skill by English teachers, and a complete match of ATP data for the cases to be studied.

A final study of interest would be an investigation into the kind of work that students of history and historians actually do. In employment settings, this kind of study is termed a job analysis. The various types of work involved in a given job are listed, and the amount of time spent on each type is estimated. The results of the job analysis are then used to weight various portions of examinations so that the job is adequately represented by the examination. For the history examinations used in the present study, one might determine, for example, to what degree work in the field of history requires a broad knowledge of history and to what degree it requires a facility in working with historical documents. The results of such a study might be used to determine the relative weights assigned to the objective and free-response portions of the AP examinations.

REFERENCES

Bleistein, C. A., R. B. Flesher, and B. Maneckshana. 1987. *Test Analysis: College Board Advanced Placement Examination, American History, 3IBP.* Report No. SR-87–33. Princeton, NJ: Educational Testing Service.

Bleistein, C. A., D. McLean, and R. B. Flesher. 1986. *Test Analysis: College Board Advanced Placement Examination, European History, 3IBP.* Report No. SR-86–151. Princeton, NJ: Educational Testing Service.

Breland, H. M. 1977. *Group Comparisons for the Test of Standard Written English.* College Entrance Examination Board Research and Development Report RDR 77–78, No. 1 (Educational Testing Service Research Bulletin RB-77–15). Princeton, NJ.

Breland, H. M., and P. A. Griswold. 1982. "Use of a Performance Test as a Criterion in a Differential Validity Study." *Journal of Educational Psychology 74* (5):713–721.

Bridgeman, B., and C. Lewis. 1990. *A Comparison of the Predictive Validity of Advanced Placement Essay and Multiple-Choice Scores.* Unpublished draft report. Princeton, NJ: Educational Testing Service.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2d ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., and P. Cohen. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2d ed.). Hillsdale, NJ: Erlbaum.

Droege, R. C. 1967. "Sex Differences in Aptitude Maturation during High School." *Journal of Counseling Psychology 14*:407–411.

Holt, T. 1990. *Thinking Historically: Narrative, Imagination, Understanding.* New York: College Entrance Examination Board.

Klein, S. P. 1989. *Does Performance Testing on the Bar Examination Reduce Differences in Scores among Sex and Racial Groups?* Unpublished manuscript.

Lord, F. M., and M. R. Novick. 1968. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley Publishing Co.

BEST COPY AVAILABLE

Maccoby, E. E., and C. N. Jacklin. 1974. *The Psychology of Sex Differences* (vol. 1). Stanford, CA: Stanford University Press.

Mazzeo, J., A. Schmitt, and C. Bleistein. 1990. *Exploratory Analyses of Some Possible Causes for the Discrepancies in Gender Differences on Multiple-Choice and Free-Response Sections of the Advanced Placement Examinations*. Draft Report. Princeton, NJ: Educational Testing Service.

Messick, S. 1989. "Validity." In *Educational Measurement*, edited by R. L. Linn. New York: Macmillan.

Petersen, N., and S. L. Livingston. 1982. *English Composition Test with Essay: A Descriptive Study of the Relationship between Essay and Objective Scores by Ethnic Group and Sex*. ETS Statistical Report No. SR-82–96. Princeton, NJ: Educational Testing Service.

Sincoff, J. B., and R. J. Sternberg. 1987. "Two Faces of Verbal Ability." *Intelligence 11*:263–276.

Sincoff, J. B., and R. J. Sternberg. 1988. "Development of Verbal Fluency Abilities and Strategies in Elementary-School-Age Children." *Developmental Psychology 24*(5):646–653.

Stevenson, H. W., R. E. Klein, G. A. Hale, and L. K. Miller. 1968. "Solution of Anagrams: A Developmental Study." *Child Development 39*:905–912.

APPENDIX A. ENGLISH RATING PROCEDURES

The rating form used for the English ratings is reproduced as Form A. Two types of English ratings were made: a holistic rating of the overall quality of each of the responses in the free-response test booklet and an English mechanics rating that applied to both responses collectively. A total of 18 English teachers made the ratings in a one-day conference session directed by Educational Testing Service humanities staff experienced with essay readings. Half of the English teachers were male and half were female. The readers were provided with no information about the history responses they rated. They were not given the history question that prompted the response, but were instructed to score the responses from an English quality perspective only. Each free-response booklet was rated by two different English teachers.

Form A. Description of the English Ratings

A. *Holistic Rating*. This is a rating on a scale from 1 (low) to 6 (high) of the *overall quality* of the discourse represented in the response. Organization, logical development, and expression should be the dominant factors in the holistic rating. This rating is made separately for each of the two responses in the booklet.

B. *Mechanics*. This rating is on a five-point scale, as indicated below, and is based on *both* responses in the booklet.

Rating	*Description of Response Qualifying for this Rating*
5	Almost no errors in usage, punctuation, agreement, spelling. Clear command of grammar, choice of words, sentence structure.
4	Few errors in usage, punctuation, agreement, spelling. General command of grammar, choice of words, sentence structure.
3	Errors in usage, punctuation, agreement, spelling. Uneven command of grammar, choice of words, sentence structure.
2	Consistent errors in usage, punctuation, agreement, spelling. Little command of grammar, choice of words, sentence structure.
1	Virtually no command of grammatical and syntactical conventions or no response.

APPENDIX B. HANDWRITING, NEATNESS, AND LENGTH ESTIMATION PROCEDURES

The rating form used for the handwriting and neatness ratings and for response length estimation appears as Form B. These ratings and estimations were made by clerical staff. Each free-response booklet was rated twice, with one rating made by a male and the other by a female. The handwriting and neatness ratings were made for both responses in the booklet taken collectively. The length estimation was made for the Part A and Part B responses separately.

Form B. Other Ratings

A. *Handwriting Quality, Legibility* (based on both responses in the booklet)

Rating	*Description of Response Qualifying for this Rating*
5	Superior, even majestic, and easy to read.
4	Good handwriting and easy to read.
3	Readable, but not pretty.
2	Difficult to read in places.
1	Very difficult to read generally.

B. *Neatness* (based on both responses in the booklet)

Rating	*Description of Response Qualifying for this Rating*
5	Extremely neat, no cross-outs, no erasures or marks in margin.
4	Neat with occasional words crossed out.
3	Not neat, but not messy.
2	A little messy.
1	Extremely messy.

C. *Length*

This is an estimate of the length of the response in words based on a count of lines written (and not crossed out) and the size of the writing. It will serve as a measure of fluency. This estimate is made separately for each response in the booklet.

APPENDIX C. U.S. HISTORY CONTENT RATING PROCEDURES

Eight different rating forms were used for the U.S. History content ratings. These forms, which appear on the following pages as Forms C-1 through C-8, were developed with the assistance of AP U.S. History staff at Educational Testing Service. In the 1986 AP U.S. History Examination, Question 1 (Part A) was a document-based question and Part B allowed a choice among five different questions (Questions 2 through 6). Within Question 6, a choice could be made among three different options.

Two graduate history students, one male and one female, used the rating forms to generate scores for main points made, supporting evidence used, factual errors made, and responsiveness. These first three ratings were combined into a single score, historical content.

Because of the labor-intensiveness of these ratings, they were conducted in a manner similar to that used for the regular AP free-response scoring: The two responses in each booklet were rated by different readers and the ratings summed. For the present study, the male graduate student read one question and the female graduate student read the other question. Each student read a random half of each of the two responses.

Form C-1

AP # _____ U.S. History Reader _____

1. The 1920s were a period of tension between new and changing attitudes on the one hand and traditional values and nostalgia on the other. What led to the tension between old and new AND in what ways was the tension manifested?

MAIN POINTS:

What led to?
- ____ World War I
- ____ A shift away from traditional social beliefs
- ____ New technologies and inventions (movies, cars, etc.)
- ____ Advances in science
- ____ Urbanization
- ____ Other reasonable cause
- ____ Totally unreasonable cause

In what ways (How?)
- ____ materialism, consumerism, advertising
- ____ Prohibition
- ____ changing attitudes toward religion
- ____ modern science vs. Protestant fundamentalism
- ____ rural values (home, community) vs. urban
- ____ racism
- ____ nativism
- ____ clinging to the traditional for some
- ____ the new morality
- ____ changing status of women
- ____ conformity, loss of individualism
- ____ other appropriate example
- ____ totally inappropriate example

SUPPORTING EVIDENCE:

	Used Correctly	Used Incorrectly
A. *Babbitt*	____	____
B. "The Bridge"	____	____
C. Evolution	____	____
D. Klan	____	____
E. Negro artist	____	____
F. Lindbergh	____	____
G. "Women Smokers"	____	____
H. Divorce rate	____	____
I. Aimee McPherson	____	____

External references: Number used correctly ____
Number used incorrectly ____

FACTUAL ERRORS:_____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form C-2

AP # _____ U.S. History Reader _____

2. "Despite the view of some historians that the conflict between Great Britain and its thirteen North American colonies was economic in origin, in fact the American Revolution had its roots in politics and other areas of American life."
   Assess the validity of this statement.

MAIN POINTS:

_____ End of salutary neglect
_____ Mercantilism was a principal cause of war (economic)
_____ Cause of war was a series of laws enacted (both economic and political)
_____ But other British colonies subjected to similar laws did not revolt
_____ Growth of republican thinking (Enlightenment thought)
_____ Colonies were prospering and wanted to expand westward
_____ Colonies no longer needed British protection after defeat of French
_____ Representation was the real issue, not taxation (political)
_____ Disregard by British of individual rights (religion, meetings, trials)
_____ America had grown away from Britain (non-British immigrants, etc.)
_____ Americans perceived themselves as different (ideology)
_____ Other reasonable cause of war
_____ Incorrect cause of war

SUPPORTING EVIDENCE:

_____ Proclamation of 1763 (closed West to settlement)
_____ Revenue Act of 1764 (mercantilism enforced) to raise money for French and Indian Wars
_____ Quartering Act (troops stationed in private homes)
_____ Stamp Act (enacted in 1765, repealed in 1766)
_____ Townshend Acts (duties to pay local officials and soldiers)
_____ Writs of Assistance
_____ Vice-Admiralty Courts
_____ Boston Massacre (due to bad relations with British soldiers)
_____ Committees of Correspondence
_____ Boycott of British goods
_____ Tea Act/ Boston Tea Party
_____ Intolerable Acts
_____ Declaratory Act
_____ First Continental Congress (import-export boycott agreement)
_____ Restraining Act (harsh British reaction to boycott)
_____ Common Sense/Thomas Paine
_____ John Locke, Enlightenment thinking
_____ Virtual representation
_____ Other reasonable evidence
_____ Incorrect supporting evidence

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form C-3

AP # _____ U.S. History Reader_____

3. "In the first half of the nineteenth century, the American cultural and intellectual community contributed to the development of a distinctive American national consciousness."
   Assess the validity of this statement.

MAIN POINTS:

_____ True, this period was an American Renaissance in literature, art, etc.
_____ True, before this period Americans had turned to Europe for culture
_____ True, before this period Europeans had no interest in American culture
_____ True, it was done by creating and popularizing things American
_____ True, America is different from Europe, we are innocent and uncorrupted
_____ True, America stresses the responsibility of the individual

SUPPORTING EVIDENCE:

Literature:
_____ Washington Irving, stories set in American locations
_____ James Fenimore Cooper, stories set in American locations
_____ Natty Bumppo character embodied American self-reliance
_____ Daniel Boone, the self-reliant, uncorrupted frontiersman
_____ George Washington cherry tree myth, Weems's biography
_____ Transcendentalist movement (belief in man)
_____ Ralph Waldo Emerson (worship comes from within man)
_____ Henry David Thoreau (worship through observing nature)
_____ "The American Scholar" (Emerson)
_____ *Walden*, "Civil Disobedience" (Thoreau)
_____ Individual knows what is right
_____ Walt Whitman
_____ John Greenleaf Whittier
_____ Louisa May Alcott, strong moral values important

Art:
_____ Hudson River School
_____ Thomas Cole
_____ Asher Durand
_____ Landscapes, American wilderness paintings

Other:
_____ Noah Webster, American Dictionary of English Language (says America different from England)
_____ Horace Mann, established schools to educate Americans here
_____ Free education so more than aristocracy would be educated
_____ Reform movements of 1840s, notion of perfectibility
_____ Second Great Awakening of 1820s, religious fervor
_____ Mormonism born in New England
_____ Evangelical Protestantism began (traveling clergymen)
_____ "Era of Good Feelings" under James Madison
_____ Following War of 1812, defeat of British, times better
_____ Nationalism began to grow during this period
_____ Other reasonable evidence

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form C-4

AP # _____ U.S. History Reader_____

4. "Both the Mexican War and the Spanish-American War were premeditated affairs resulting from deliberately calculated schemes of robbery on the part of a superior power against weak and defenseless neighbors."
   Assess the validity of this statement.

MAIN POINTS:

_____ Both conflicts clearly showed American imperialism
_____ Both conflicts had a long history, so not schemes
_____ Neither Mexico nor Spain was defenseless
_____ Both wars followed acts of violence against U.S.
_____ People in territories involved in revolution with their governments
_____ U.S. paid for territories obtained
_____ Ongoing rebellion in Cuba was hurting American investors in Cuba
_____ U.S. provoked Mexican War to gain California ports
_____ U.S. had always wanted Cuba, all Europe grabbing colonies
_____ An unstable Cuba threatened proposed Panama canal
_____ Other

SUPPORTING EVIDENCE:

Mexican War:

_____ American cavalry attacked by Mexicans between Rio Grande and Nueces
_____ Texas admitted as state in 1845, Mexico broke relations with U.S.
_____ Both Texas and California unhappy with Mexican rule
_____ Mexico in revolution, country out of control, owed U.S. money
_____ U.S. had offered to buy territories desired for $30 million
_____ U.S. eventually paid $15 million for territories
_____ Manifest Destiny
_____ Treaty of Guadalupe Hidalgo
_____ Zachary Taylor elected president in 1848
_____ John Quincy Adams and Whig opponents of war declared it immoral
_____ Thoreau refused to pay taxes for war
_____ Other

Spanish-American War:

_____ Sinking of battleship *Maine* in Havana, loss of 260 lives
_____ Cuban people in revolt against cruel Spanish governor Weyler y Nicolau
_____ American people sympathetic with oppressed people of Cuba
_____ Hearst newspapers inflamed public opinion, jingoism
_____ U.S. used Cuba as excuse to seize Philippines, Guam, and Puerto Rico
_____ U.S. paid $20 million for property in Philippines
_____ Treaty of Paris
_____ Ostend Manifesto
_____ Other

FACTUAL ERRORS:_____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form C-5

AP # _____ U.S. History Reader _____

5. Andrew Carnegie has been viewed by some historians as the "prime representative of the industrial age" and by others as "an industrial leader atypical of the period."
   Assess the validity of these views.

MAIN POINTS:

Typical:
- _____ Self-made man
- _____ Robber baron
- _____ Industrial statesman
- _____ Ruthless in business dealings
- _____ Used his genius, eye for opportunity, to build his business
- _____ Great organizational skills
- _____ Created plant model for modern steel industry
- _____ Believed in laissez-faire government
- _____ Helped bring order out of chaos
- _____ Social Darwinist—the fittest survive
- _____ Philosophy of Herbert Spencer
- _____ Other

Atypical:
- _____ Philanthropist
- _____ Humane, had a conscience
- _____ Became a strong opponent of imperialism
- _____ Immigrant (from Scotland)
- _____ Other

SUPPORTING EVIDENCE:

Typical:
- _____ Built U.S. Steel
- _____ Integrated all stages of production in one mill
- _____ Used Bessemer process initially
- _____ Vanderbilt (railroads)
- _____ Rockefeller (oil)
- _____ McCormick (farm implements)
- _____ Swift (meat packing)
- _____ Rags-to-riches myth
- _____ Other

Atypical:
- _____ Gave away three-fourths of his wealth
- _____ Did not practice conspicuous consumption like Gould, others
- _____ "Gospel of Wealth" (article by Carnegie)
- _____ Established numerous libraries
- _____ Really poor, not like Rockefeller
- _____ Other

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form C-6

AP # _____ U.S. History Reader _____

6a. "Reform movements of the twentieth century have shown continuity in their goals and strategies." Assess the validity of this statement for ONE of the following pairs of reform movements.

Progressivism and the New Deal
Woman's suffrage and post–Second World War feminism
The New Deal and the Great Society

MAIN POINTS:

_____ Both had underlying goal of preserving democracy and capitalism
_____ Specific goals differed
_____ Both wanted to eliminate abuses
_____ Both wanted to preserve the middle class
_____ Both used same strategy—legislation
_____ Progressivism sought to end corruption in government, make it more responsive
_____ Progressivism sought to reform the economy
_____ Progressivism sought to make economy fair, but left responsibility to individual
_____ New Deal took responsibility for economy and welfare of people
_____ New Deal sought to end Depression and help people survive
_____ Progressivism and New Deal different in terms of individual responsibility
_____ Other

SUPPORTING EVIDENCE:

Progressivism:

_____ Sherman Anti-Trust Act (break up trusts)
_____ Clayton Anti-Trust Act
_____ Federal Reserve System (bank regulation)
_____ Food and Drug Acts, meat inspection
_____ Child labor laws
_____ Maximum working hours
_____ Direct election of senators
_____ City manager form of government
_____ Other

New Deal:

_____ NRA
_____ CCC
_____ NLRB
_____ SEC
_____ TVA, government ownership
_____ Wagner Act (sought to give labor power to match big business)
_____ Social Security, government will take care of little people
_____ Other

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form C-7

AP # _____ U.S. History Reader _____

6b. "Reform movements of the twentieth century have shown continuity in their goals and strategies." Assess the validity of this statement for ONE of the following pairs of reform movements.

Progressivism and the New Deal
Woman's suffrage and post–Second World War feminism
The New Deal and the Great Society

MAIN POINTS:

_____ Suffrage movement sought the right to vote
_____ Both sought to have women recognized as equals
_____ Feminism sought equality in workplace
_____ Feminism sought equality in political system
_____ Feminism sought equality in home
_____ Strategies similar: both sought allies to get legislation passed
_____ Both used marches and demonstrations to publicize demands
_____ Other

SUPPORTING EVIDENCE:

_____ 19th Amendment
_____ Susan B. Anthony
_____ Carne Chapman Catt—NAWSA
_____ Alice Paul—militant National Women's Party
_____ Cady Stanton
_____ WCTU, anti-smoking
_____ Gloria Steinem, *Ms. Magazine*
_____ Betty Friedan, *The Feminine Mystique*
_____ Geraldine Ferraro, example of women in politics
_____ WW II, women in work force
_____ NOW, professional women, founded 1966
_____ 1963 Presidential Commission on the Status of Women
_____ 1964 Civil Rights Act (Title VII)
_____ Title IX, 1972
_____ Demands for child care, equal pay, abortion reform
_____ *Roe* v. *Wade* (abortion reform)
_____ Formerly all-male colleges admit women in 1970s
_____ Affirmative action
_____ ERA
_____ Other

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form C-8

AP # _____ U.S. History Reader _____

6c. "Reform movements of the twentieth century have shown continuity in their goals and strategies." Assess the validity of this statement for ONE of the following pairs of reform movements.

Progressivism and the New Deal
Woman's suffrage and post–Second World War feminism
The New Deal and the Great Society

MAIN POINTS:

_____ Goals and strategies of both similar
_____ Both sought to use government to make reforms through legislation
_____ Both sought to take care of people at bottom
_____ Neither stressed individual responsibility for economy
_____ Stress on regulation and management as opposed to free enterprise
_____ One difference: the Great Society stressed civil rights legislation, voting rights
_____ Other

SUPPORTING EVIDENCE:

New Deal:

_____ NRA
_____ CCC
_____ NLRB
_____ SEC
_____ TVA, government ownership
_____ Wagner Act (sought to give labor power to match big business)
_____ Social Security
_____ Other

Great Society:

_____ War on Poverty
_____ Head Start
_____ Job Corps
_____ Medicare
_____ Federal Aid to Education
_____ Voting Rights Act
_____ Cabinet posts for Housing, Urban Affairs, Transportation
_____ Other

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

APPENDIX D. EUROPEAN HISTORY CONTENT RATING PROCEDURES

Seven different rating forms were used for the European History content ratings. These forms, which appear as Forms D-1 through D-7, were developed with the assistance of a consultant in European history who served as a regular Reader for AP European History Examinations. In the 1986 AP European History Examination, Part A was a document-based question (Question 1) and Part B allowed a choice among six different questions (Questions 2 through 7).

Two graduate history students, one male and one female, used the rating forms to generate scores for main points made, supporting evidence used, factual errors made, and responsiveness. The first three ratings were combined into a single score, historical content.

Because of the labor-intensiveness of these ratings, they were conducted in a manner similar to that used for the regular AP free-response scoring: The two responses in each booklet were rated by different readers and the ratings summed. For the present study, the male graduate student read one question and the female graduate student read the other question. Each student read a random half of each of the two responses.

Form D-1

AP # _____ European History Reader # _____

1. Analyze the pressures on Great Britain's Liberal government during the Sudan Crisis (1884–1885), and explain why the government acted as it did. Use documents taking into account their source and the point of view of the author. You may also use other historical facts.

MAIN POINTS:

Pressures:
- _____ Liberal government at risk, would be responsible
- _____ Conservatives were using incident to damage Liberals
- _____ British interests and citizens in Near East at risk
- _____ Public outcry to rescue the popular Gordon
- _____ To maintain England's reputation among her colonies
- _____ Other correct pressure
- _____ Totally incorrect pressure suggested

Why government acted as it did:
- _____ Both Liberals and Conservatives divided on issue
- _____ Gordon had been instructed to evacuate, but did not
- _____ Gladstone didn't want an immoral war of conquest
- _____ Gladstone didn't want a religious war
- _____ The distance to Khartoum made military action difficult
- _____ Egypt already resentful of British occupation
- _____ Other correct reason
- _____ Totally incorrect reason suggested

SUPPORTING EVIDENCE:	Source/ Significance	Point of View
Doc. #1. Map	_____	_____
Doc. #2. Gordon starts out	_____	_____
Doc. #3. Marquess of Salisbury	_____	_____
Doc. #4. Sir George Campbell	_____	_____
Doc. #5. Sir Wilfred Lawson	_____	_____
Doc. #6. Gordon isolated	_____	_____
Doc. #7. Dilke memoirs	_____	_____
Doc. #8. Gladstone memo	_____	_____
Doc. #9. Hicks-Beach	_____	_____
Doc. #10. Gladstone	_____	_____
Doc. #11. Sir Evelyn Baring	_____	_____
Doc. #12. *Punch* cartoon	_____	_____
Doc. #13. Gladstone	_____	_____

External References: # used correctly _____
used incorrectly _____

FACTUAL ERRORS:_____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form D-2

AP # _____ European History Reader_____

2. "The Reformation was a rejection of the secular spirit of the Italian Renaissance." Defend or refute this statement using specific examples from sixteenth-century Europe.

MAIN POINTS:

_____ Rejection in sense of changes in architecture, art, science, and politics
_____ Reformation was different in different countries
_____ Reformation was social, economic, and political—not only religious
_____ Reformation in some countries was a rejection of Catholicism
_____ Reformation rejected idea that all life should be aimed toward heaven
_____ Renaissance was not *completely* secular
_____ Renaissance emphasized secular *pleasures*, but did not reject church
_____ Despite the different ideals of the two, both were humanist movements
_____ Both the Renaissance and the Reformation gave individuals more freedom
_____ Changing social status of the middle class
_____ Rise of literacy
_____ Transfer of power from church to territorial princes and to the state
_____ Other main point

SUPPORTING EVIDENCE:

_____ Martin Luther
_____ The Ninety-five Theses
_____ John Calvin
_____ Henry VIII, Elizabeth I
_____ Selling of indulgences
_____ Simony, pluralism, nepotism
_____ "Quattrocento"
_____ The Medici family
_____ Counter-Reformation
_____ Society of Jesus
_____ Council of Trent
_____ Inquisition
_____ Branding of heretics
_____ Galileo/Copernicus
_____ City-states governed by oligarchs in Renaissance
_____ Michelangelo
_____ Leonardo da Vinci
_____ Other evidence

FACTUAL ERRORS: _____

RESPONSIVENESS

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form D-3

AP # _____ European History Reader _____

3. In the seventeenth century, how did England and the Dutch republic compete successfully with France and Spain for control of overseas territory and trade?

MAIN POINTS (How did they compete?):

_____ They founded great joint stock companies, protected by governments
_____ They developed sophisticated banking systems
_____ They encouraged shipbuilding
_____ They took advantage of financial weaknesses of France and Spain
_____ They put their energies primarily into trade, not war
_____ They allowed their colonies some independence (unlike Spain)
_____ They used their colonies primarily for raw materials (not gold)
_____ Protestant religion favored enterprise and thrift in both nations
_____ Dutch granted some religious toleration
_____ Other ways Dutch and English competed

SUPPORTING EVIDENCE:

_____ Dutch East India Company (founded 1602)
_____ Dutch West India Company (founded 1621)
_____ British East India Company
_____ Bank of Amsterdam
_____ Bank of England
_____ Wars of Religion (France)
_____ Thirty Years War (France and Spain both involved)
_____ Wars of Louix XIV of France
_____ British Navigation Acts protected interests
_____ British naval power protected interests
_____ Cromwell's mid-century naval reforms
_____ Decline in Spanish leadership (Philip III, Charles II)
_____ Decline of Bourbons and Hapsburgs
_____ Decline in yields from Spanish silver mines
_____ Mercantile groups in both England and the Dutch republic
_____ France's colonies good only for fur and timber
_____ Spain used its colonies only for silver and gold
_____ Other evidence

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form D-4

AP # _____ European History Reader _____

4. To what extent and in what ways was the French Revolution during the period 1789 through the Reign of Terror (1794) an attempt to create a government based on Enlightenment ideals?

MAIN POINTS:

_____ The Enlightenment produced a generation of people critical of governments
_____ The Enlightenment was a very strong influence on the French Revolution
_____ The Enlightenment was based on a natural rights philosophy and rejected autocracy
_____ The objective of the French Revolution was to establish a republic based on reason
_____ The ideals of the French Revolution borrowed heavily from Rousseau and Locke
_____ The Enlightenment grew and flourished in France and influenced thinking
_____ The "general will" of the people heavily influenced the French Revolution
_____ The French Revolution embraced the right of people to revolt and establish their own government
_____ The French Revolution embraced secularism and made the Church subservient to the state
_____ The revolutionaries admired the English system of government
_____ Of special importance in new government was the balance of powers
_____ Nationalism and the Reign of Terror
_____ Terror was distortion of Enlightenment ideals, but agreed with Social Contract
_____ Other main point

SUPPORTING EVIDENCE:

_____ John Locke/*Two Treatises on Government*
_____ Jean Jacques Rousseau/*The Social Contract*
_____ Montesquieu/*The Spirit of the Laws*
_____ Voltaire
_____ Hume/Gibbon/Kant/Franklin/Jefferson
_____ American Revolution
_____ *The Declaration of the Rights of Man*
_____ Tennis Court Oath
_____ Night of August 4, 1789
_____ Committee of Public Safety
_____ De-Christianization
_____ Cult of Reason
_____ Constitutional Monarchy of 1791
_____ Constitution of 1793
_____ Voltaire's enlightened absolutism/inconsistency with French Revolution
_____ Age of Reason
_____ Philosophes
_____ Robespierre
_____ Louis XVI
_____ Marie Antoinette
_____ Sans-culottes
_____ Bourgeoisie
_____ Other evidence

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form D-5

AP # _____ European History Reader _____

5. Evaluate the effectiveness of collective responses by workers to industrialization in Western Europe during the course of the nineteenth century.

MAIN POINTS:

_____ The tactics of collective responses differed in different countries
_____ The effectiveness of these responses differed in different countries
_____ Industrialization and collective responses began first in England
_____ The early tactics in England were bread riots and were not successful
_____ Later tactics in England were political and were very successful
_____ In France, tactics were very violent and not effective
_____ In France, there was no faith that government would reform itself
_____ In France, trade unions were eventually legalized by government
_____ In Germany, collective responses began later, and were easily put down
_____ In Germany, to gain popular support, reforms were instituted
_____ In summary, collective responses were most effective in England
_____ France caught up eventually with England, however
_____ In France and Germany, reforms were initiated by the governments
_____ Legislative reform through governments was most effective means
_____ Marxist theories developed during this period but not practiced
_____ General broadening of suffrage in Britain, France, and Germany
_____ Other main point

SUPPORTING EVIDENCE:

_____ Peterloo Massacre
_____ Six Acts
_____ Reform Bill of 1832
_____ Reform Bill of 1867
_____ Chartism/Chartists
_____ Anti-Corn Law League (1839)
_____ Repeal of the Corn Laws (1846)
_____ Paris Commune
_____ Napoleon III
_____ Bismarck
_____ Disraeli
_____ Legalization of unions in Britain (1824)
_____ Ten Hours Act in Britain (1847)
_____ Right to strike won in Britain in 1870s
_____ Utopian socialists in France, e.g., Proudhon, Fourier, Blanc
_____ French worker groups revolt in 1830 and 1848
_____ Le Chapelier Law (1791) repealed under Napoleon III
_____ Union rights extended under Third Republic
_____ Bismarck's anti-socialist laws (1878)
_____ Social Democratic Party
_____ *Communist Manifesto* (1848)
_____ First and Second International
_____ Eduard Bernstein/*Evolutionary Socialism*
_____ German Trade Congress
_____ Other evidence

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

38

Form D-6

AP # _____ European History Reader _____

6. To what extent and in what ways has twentieth-century physics challenged the Newtonian view of the universe and society?

MAIN POINTS:

_____ Twentieth-century physics completely challenges the Newtonian view
_____ Eighteenth-century man viewed the universe as guided by natural laws
_____ Man and society were rational things ruled by natural laws
_____ Enlightenment thinking stressed observation and reason
_____ Scientific Revolution based on observation and rational evaluation
_____ Twentieth-century physics showed that all events could not be observed
_____ Seemingly solid things were shown to consist of atoms that moved
_____ The micro world of twentieth-century physics was beyond human sight and touch
_____ The macro world of twentieth-century physics questioned the reality of time/light
_____ Twentieth-century physics replaced Newton's certainties with probabilities
_____ Application of science to political theory
_____ Application of science to religious thought
_____ Application of science to artistic movements
_____ Application of science to psychology
_____ Other main point

SUPPORTING EVIDENCE:

_____ Copernicus
_____ Newton/law of gravity
_____ Newton/synthesis of earthly and cosmic laws of motion
_____ Newton/calculus
_____ Einstein/theory of relativity
_____ Kepler/theory of elliptical orbits
_____ Galileo/telescope
_____ Alexander Pope
_____ Locke/*Second Treatise on Civil Government*
_____ Deism
_____ Planck/theories undermining observation as way to truth
_____ Heisenberg/principle of uncertainty
_____ Freud
_____ Dada/Futurists/Surrealists
_____ The heliocentric universe
_____ Philosophes
_____ Inductive reasoning
_____ Speed of planets in relation to sun
_____ Other evidence

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

Form D-7

AP # _____ European History Reader _____

7. Assess the strengths and weaknesses of the economic revival of Western Europe between 1945 and 1970.

MAIN POINTS:

_____ After World War II, Western Europe was in economic ruin
_____ Because of this economic ruin, it was feared ripe for communism
_____ U.S. dropped war debts
_____ U.S. sent large amounts of financial aid to Western European countries
_____ There was an amazing economic recovery between 1945 and 1970
_____ But financial aid created a dependency on U.S.
_____ Financial aid also isolated Communist countries and created Cold War
_____ International agreements among Western European countries helped also in recovery
_____ Western European nations also became dependent upon each other
_____ Thus economic integration promoted spread of recessions in Western Europe
_____ Conflict between technology and labor unions also weakened recovery
_____ Recovery was strong economically but weak politically and otherwise
_____ Western Europe still dependent on Middle East for oil
_____ All countries of Western Europe did not prosper as much as West Germany
_____ Inflation in the 1960s resulted in welfare states in some countries
_____ Competition from U.S. weakened recovery
_____ Other main point

SUPPORTING EVIDENCE:

_____ Marshall Plan (1947)
_____ International Monetary Fund (IMF)
_____ European Economic Community (EEC)
_____ North Atlantic Treaty Organization (NATO)
_____ Confederation of Steel and Iron Industries
_____ European Coal and Steel Community (1950)
_____ Internal tariffs eliminated (1968)
_____ Production in 1952 was 25% greater than pre–World War II levels
_____ Production in 1963 was 2.5 times pre–World War II levels
_____ Ludwig Erhard, West German economic planner
_____ Jean Monnet, French economic planner
_____ British economic growth relatively slower during recovery
_____ American multinational companies spread throughout Western European economies

FACTUAL ERRORS: _____

RESPONSIVENESS:

4 Fully responsive to all aspects of questions asked
3 Partially responsive
2 Half-hearted response
1 No response, or totally inappropriate response

40

APPENDIX E. CORRELATION TABLES

Tables E-1 through E-11 show the correlations between the variables for the two history examinations.

Table E-1. U.S. History Correlations, Full Sample

	FR	*OBJ*	*ENG*	*HAND*	*NEAT*	*WORDS*
			Total Sample (N = 398)			
FR	1.00					
OBJ	0.49	1.00				
ENG	0.50	0.35	1.00			
HAND	0.08	−0.14	0.22	1.00		
NEAT	0.00	−0.12	0.07	0.38	1.00	
WORDS	0.57	0.25	0.44	0.05	−0.01	1.00
			Males (N = 199)			
FR	1.00					
OBJ	0.48	1.00				
ENG	0.53	0.37	1.00			
HAND	0.11	−0.07	0.24	1.00		
NEAT	0.01	−0.14	0.03	0.38	1.00	
WORDS	0.61	0.32	0.48	0.12	0.05	1.00
			Females (N = 199)			
FR	1.00					
OBJ	0.54	1.00				
ENG	0.49	0.40	1.00			
HAND	0.02	−0.08	0.12	1.00		
NEAT	−0.03	−0.04	0.06	0.31	1.00	
WORDS	0.52	0.21	0.41	−0.12	−0.12	1.00

Abbreviations: FR — AP free-response score
OBJ — AP objective score
ENG — English score
HAND — Handwriting rating
NEAT — Neatness rating
WORDS — Words written

Table E-2. European History Correlations, Full Sample

	FR	*OBJ*	*ENG*	*HAND*	*NEAT*	*WORDS*
			Total Sample (N = 397)			
FR	1.00					
OBJ	0.42	1.00				
ENG	0.57	0.38	1.00			
HAND	0.10	−0.03	0.21	1.00		
NEAT	0.16	−0.08	0.25	0.51	1.00	
WORDS	0.61	0.26	0.58	0.04	0.15	1.00
			Male (N = 198)			
FR	1.00					
OBJ	0.46	1.00				
ENG	0.61	0.42	1.00			
HAND	0.03	0.05	0.22	1.00		
NEAT	0.10	−0.02	0.23	0.48	1.00	
WORDS	0.64	0.25	0.62	−0.02	0.11	1.00
			Females (N = 199)			
FR	1.00					
OBJ	0.41	1.00				
ENG	0.53	0.37	1.00			
HAND	0.19	−0.07	0.18	1.00		
NEAT	0.24	−0.12	0.27	0.53	1.00	
WORDS	0.57	0.25	0.54	0.14	0.22	1.00

Abbreviations: FR — AP free-response score
OBJ — AP objective score
ENG — English score
HAND — Handwriting rating
NEAT — Neatness rating
WORDS — Words written

Table E-3. U.S. History Correlations, Full Sample—Part A

	FR.A	*OBJ*	*HOLA*	*MECH*	*WORDS.A*
		Total Sample (N = 398)			
FR.A	1.00				
OBJ	0.38	1.00			
HOLA	0.36	0.19	1.00		
MECH	0.40	0.34	0.54	1.00	
WORDS.A	0.50	0.14	0.33	0.30	1.00
		Males (N = 199)			
FR.A	1.00				
OBJ	0.38	1.00			
HOLA	0.42	0.21	1.00		
MECH	0.36	0.35	0.61	1.00	
WORDS.A	0.55	0.25	0.46	0.33	1.00
		Females (N = 199)			
FR.A	1.00				
OBJ	0.38	1.00			
HOLA	0.42	0.21	1.00		
MECH	0.36	0.35	0.61	1.00	
WORDS.A	0.55	0.25	0.46	0.33	1.00

Abbreviations: FR.A — AP free-response score, Part A
OBJ — AP objective score
HOLA — Holistic English score, Part A
MECH — English mechanics score
WORDS.A — Words written, Part A

Table E-4. European History Correlations, Full Sample—Part A

	FR.A	*OBJ*	*HOLA*	*MECH*	*HAND*	*NEAT*	*WORDS.A*
			Total (N = 397)				
FR.A	1.00						
OBJ	0.28	1.00					
HOLA	0.52	0.33	1.00				
MECH	0.43	0.33	0.83	1.00			
HAND	0.10	−0.03	0.20	0.23	1.00		
NEAT	0.16	−0.08	0.24	0.25	0.51	1.00	
WORDS.A	0.55	0.16	0.58	0.39	0.08	0.12	1.00
			Male (N = 198)				
FR.A	1.00						
OBJ	0.30	1.00					
HOLA	0.55	0.34	1.00				
MECH	0.48	0.36	0.85	1.00			
HAND	0.02	0.05	0.21	0.25	1.00		
NEAT	0.11	−0.02	0.23	0.23	0.48	1.00	
WORDS.A	0.59	0.11	0.58	0.43	0.03	0.09	1.00
			Female (N = 199)				
FR.A	1.00						
OBJ	0.25	1.00					
HOLA	0.49	0.34	1.00				
MECH	0.38	0.32	0.80	1.00			
HAND	0.21	−0.07	0.18	0.18	1.00		
NEAT	0.22	−0.12	0.24	0.25	0.53	1.00	
WORDS.A	0.50	0.20	0.58	0.36	0.16	0.15	1.00

Abbreviations: FR.A — AP free-response score, Part A
OBJ — AP objective score
HOLA — Holistic English score, Part A
MECH — English mechanics score
HAND — Handwriting rating
NEAT — Neatness rating
WORDS, A — Words written, Part A

Table E-5. U.S. History Correlations, Full Sample—Part B

	FR.B	OBJ	HOLB	MECH	WORDS.B
		Total (N = 398)			
FR.B	1.00				
OBJ	0.42	1.00			
HOLB	0.40	0.28	1.00		
MECH	0.28	0.34	0.39	1.00	
WORDS.B	0.54	0.30	0.43	0.25	1.00
		Males (N = 199)			
FR.B	1.00				
OBJ	0.41	1.00			
HOLB	0.52	0.44	1.00		
MECH	0.33	0.35	0.78	1.00	
WORDS.B	0.55	0.32	0.56	0.27	1.00
		Females (N = 199)			
FR.B	1.00				
OBJ	0.43	1.00			
HOLB	0.40	0.26	1.00		
MECH	0.24	0.39	0.25	1.00	
WORDS.B	0.52	0.29	0.42	0.24	1.00

Abbreviations: FR.B — AP free-response score, Part B
OBJ — AP objective score
HOLB — Holistic English score, Part B
MECH — English mechanics score
WORDS, B — Words written, Part B

Table E-6. European History Correlations, Full Sample—Part B

	FR.B	OBJ	HOLB	MECH	HAND	NEAT	WORDS.B
			Total (N = 397)				
FR.B	1.00						
OBJ	0.41	1.00					
HOLB	0.52	0.43	1.00				
MECH	0.35	0.33	0.79	1.00			
HAND	0.06	−0.03	0.14	0.23	1.00		
NEAT	0.11	−0.08	0.20	0.25	0.51	1.00	
WORDS.B	0.49	0.28	0.66	0.38	−0.02	0.15	1.00
			Male (N = 198)				
FR.B	1.00						
OBJ	0.46	1.00					
HOLB	0.58	0.48	1.00				
MECH	0.37	0.36	0.79	1.00			
HAND	0.03	0.05	0.13	0.25	1.00		
NEAT	0.05	−0.02	0.17	0.23	0.48	1.00	
WORDS.B	0.54	0.32	0.71	0.43	−0.07	0.10	1.00
			Female (N = 199)				
FR.B	1.00						
OBJ	0.39	1.00					
HOLB	0.45	0.39	1.00				
MECH	0.33	0.32	0.80	1.00			
HAND	0.09	−0.07	0.15	0.18	1.00		
NEAT	0.17	−0.12	0.25	0.25	0.53	1.00	
WORDS.B	0.44	0.22	0.59	0.33	0.07	0.22	1.00

Abbreviations: FR.B — AP free-response score, Part B
OBJ — AP objective score
HOLB — Holistic English score, Part B
MECH — English mechanics score
HAND — Handwriting rating
NEAT — Neatness rating
WORDS, B — Words written, Part B

Table E-7. Correlations for Historical Content Sample (N = 300)

	OBJ	FR	CONT	MP	SE	ENG	HAND	NEAT	WORDS
				U.S. History					
OBJ	1.00								
FR	.50	1.00							
CONT	.59	.63	1.00						
MP	.46	.48	.73	1.00					
SE	.51	.56	.91	.38	1.00				
ENG	.35	.51	.44	.38	.37	1.00			
HAND	.13	.03	.00	.03	−.02	.22	1.00		
NEAT	.09	−.01	−.05	.02	−.07	.09	.39	1.00	
WORDS	.22	.54	.50	.41	.43	.43	.03	−.01	1.00
				European History					
OBJ	1.00								
FR	.43	1.00							
CONT	.47	.69	1.00						
MP	.36	.51	.71	1.00					
SE	.42	.63	.92	.37	1.00				
ENG	.38	.63	.61	.43	.56	1.00			
HAND	.01	.15	.08	.04	.08	.23	1.00		
NEAT	−.04	.23	.14	.08	.13	.30	.55	1.00	
WORDS	.28	.64	.63	.44	.58	.63	.11	.21	1.00

Abbreviations: OBJ — AP objective score
FR — AP free-response score
CONT — Historical content score
MP — Main points
SE — Supporting evidence
ENG — English score
HAND — Handwriting rating
NEAT — Neatness rating
WORDS — Words written

Table E-8. Correlations by Sex: U.S. History, Historical Content Sample

	OBJ	FR	CONT	MP	SE	ENG	HAND	NEAT	WORDS
				Male (N = 150)					
OBJ	1.00								
FR	.49	1.00							
CONT	.56	.61	1.00						
MP	.42	.42	.74	1.00					
SE	.50	.56	.90	.37	1.00				
ENG	.38	.53	.46	.38	.39	1.00			
HAND	−.08	.04	−.01	.01	−.02	.21	1.00		
NEAT	−.16	−.03	−.09	−.01	−.12	.04	.42	1.00	
WORDS	.31	.57	.54	.48	.44	.45	.05	.04	1.00
				Female (N = 150)					
OBJ	1.00								
FR	.53	1.00							
CONT	.63	.66	1.00						
MP	.52	.54	.73	1.00					
SE	.54	.56	.91	.38	1.00				
ENG	.37	.50	.43	.38	.35	1.00			
HAND	−.06	.01	.01	.04	−.02	.16	1.00		
NEAT	.03	.02	.01	.05	−.02	.11	.30	1.00	
WORDS	.15	.52	.46	.33	.42	.42	−.05	−.10	1.00

Abbreviations: OBJ — AP objective score
FR — AP free-response score
CONT — Historical content score
MP — Main points
SE — Supporting evidence
ENG — English score
HAND — Handwriting rating
NEAT — Neatness rating
WORDS — Words written

Table E-9. Correlations by Sex: European History, Historical Content Sample

	OBJ	FR	CONT	MP	SE	ENG	HAND	NEAT	WORDS
Male (N = 150)									
OBJ	1.00								
FR	.49	1.00							
CONT	.45	.70	1.00						
MP	.33	.49	.69	1.00					
SE	.40	.64	.91	.33	1.00				
ENG	.40	.67	.67	.47	.61	1.00			
HAND	.10	.05	.01	−.02	.02	.20	1.00		
NEAT	.01	.16	.08	.07	.07	.26	.53	1.00	
WORDS	.28	.68	.70	.46	.65	.69	.03	.18	1.00
Female (N = 150)									
OBJ	1.00								
FR	.39	1.00							
CONT	.51	.68	1.00						
MP	.39	.53	.73	1.00					
SE	.46	.62	.93	.42	1.00				
ENG	.35	.58	.52	.38	.49	1.00			
HAND	−.04	.25	.18	.15	.16	.29	1.00		
NEAT	−.07	.31	.21	.11	.21	.35	.57	1.00	
WORDS	.25	.61	.55	.41	.50	.56	.25	.26	1.00

Abbreviations: OBJ — AP objective score
FR — AP free-response score
CONT — Historical content score
MP — Main points
SE — Supporting evidence
ENG — English score
HAND — Handwriting rating
NEAT — Neatness rating
WORDS — Words written

Table E-10. Correlations with Matched ATP Variables for U.S. History

Variable	OBJ	SAT-V	SAT-M	READ	VOCAB	TSWE	ECT*
Total sample (N = 276)							
AP free-response score	.51	.39	.26	.39	.33	.38	.39
Part A score	.41	.32	.20	.33	.25	.37	.40
Part B score	.39	.30	.21	.29	.26	.24	.20
Males (N = 138)							
AP free-response score	.52	.33	.15	.32	.27	.28	.41
Part A score	.42	.23	.14	.22	.19	.30	.34
Part B score	.41	.29	.09	.29	.24	.15	.28
Females (N = 138)							
AP free-response score	.51	.46	.41	.48	.38	.49	.36
Part A score	.46	.42	.35	.46	.34	.43	.41
Part B score	.35	.30	.30	.28	.27	.34	.13

*For ECT matched data, sample sizes were reduced as follows:
Total sample, $N = 140$
Males, $N = 72$
Females, $N = 68$

Abbreviations: OBJ — AP objective score
SAT-V — SAT-verbal
SAT-M — SAT-mathematical
READ — SAT reading
VOCAB — SAT vocabulary
TSWE — Test of Standard Written English
ECT — English Composition Test

Table E-11. Correlations with Matched ATP Variables for European History

Variable	OBJ	SAT-V	SAT-M	READ	VOCAB	TSWE	ECT*
Total sample (N = 272)							
AP free-response score	.44	.35	.15	.34	.32	.21	.16
Part A score	.30	.30	.14	.29	.29	.14	.12
Part B score	.43	.28	.11	.28	.25	.20	.14
Males (N = 136)							
AP free-response score	.47	.37	.21	.32	.38	.25	.16
Part A score	.34	.36	.23	.31	.36	.21	.19
Part B score	.46	.27	.13	.23	.28	.21	.08
Females (N = 136)							
AP free-response score	.41	.33	.09	.37	.26	.17	.17
Part A score	.28	.25	.06	.27	.21	.07	.04
Part B score	.40	.28	.08	.33	.21	.20	.23

*For ECT matched data, sample sizes were reduced as follows:
Total sample, $N = 167$
Males, $N = 87$
Females, $N = 80$

Abbreviations: OBJ — AP objective score
SAT-V — SAT-verbal
SAT-M — SAT-mathematical
READ — SAT reading
VOCAB — SAT vocabulary
TSWE — Test of Standard Written English
ECT — English Composition Test

BEST COPY AVAILABLE

APPENDIX F. RELIABILITY ESTIMATES

Table F-1 shows the reliability estimates for the variables for the two history examinations.

Table F-1. Reliability Estimates

Variable	*U.S. History*	*European History*
AP objective score	.90*	.90*
AP free-response score	.47**	.55**
Historical content score	.59**	.62**
Holistic English score	.72**	.76**
English mechanics score	.77**	.76**
Handwriting score	.64**	.75**
Neatness score	.76**	.78**

*KR-20 estimates from 1986 test administration.

**Alpha estimates for random samples of 300 cases for each area.

The alpha estimates were made using procedures described in Lord and Novick (1968: 88–91). These estimates violate a number of assumptions, however. The estimates for the AP free-response score and for the historical content score are derived from a variance-covariance matrix developed from one reader's score on Part A and another reader's score on Part B and assume that the two free-response parts are measuring the same thing and are parallel measures. Given the differences between the two free-response parts, it seems likely that this assumption is violated. Moreover, Part B allows a choice of topics, and no reliability estimation methods have been developed where a choice is allowed. These estimates might be considered to be lower bound estimates, however.

The estimates for the holistic English score are based on ratings made by two readers working independently. Since the readers rated Parts A and B of each booklet at the same time, however, variance-covariance matrices were developed for two combinations of scores: Part A, Reader 1, plus Part B, Reader 2; and Part A, Reader 2, plus Part B, Reader 1. The estimates derived from each pair of scores were then averaged.

The estimates for the English mechanics score, the handwriting score, and the neatness score are essentially interrater reliabilities because both free responses (Part A and Part B) were rated as a unit and given one score. The variance-covariance matrices were developed from two reader scores for each of the three types of ratings.

APPENDIX G. INTERACTION ANALYSES

Tables G-1 and G-2 show the interactions between sex and predictor variables.

Table G-1. Analyses of Interactions between Sex and Covariates in the Prediction of Residualized Criterion (Predicted Free-Response Score Minus Actual Free-Response Score) for U.S. History

Variables	R-*Squared*	R-*Squared Increase*
English	.1529**	
English and sex	.1583**	.0054
English, sex, and product	.1588**	.0005
Content	.1516**	
Content and sex	.1661**	.0145*
Content, sex, and product	.1682**	.0021
All covariates	.2111**	
Covariates and sex	.2187**	.0076
Covariates, sex, and products	.2210**	.0023

*$p < .05$
**$p < .01$

Table G-2. Analyses of Interactions between Sex and Covariates in the Prediction of Residualized Criterion (Predicted Free-Response Score Minus Actual Free-Response Score) for European History

Variables	R-*Squared*	R-*Squared Increase*
English	.2693**	
English and sex	.2852**	.0159*
English, sex, and product	.2852**	.0000
Content	.2956**	
Content and sex	.3123**	.0167**
Content, sex, and product	.3123**	.0000
Neatness	.0775**	
Neatness and sex	.0868**	.0093
Neatness, sex, and product	.0976	.0108
All covariates	.3736**	
Covariates and sex	.3879**	.0143**
Covariates, sex, and products	.3937**	.0058

*$p < .05$
**$p < .01$